VARIABLE SELECTION AND STRUCTURAL DISCOVERY IN JOINT MODELS OF LONGITUDINAL AND SURVIVAL DATA

Zangdong He

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the Department of Biostatistics,
Indiana University

December, 2014

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Wanzhu Tu, Ph.D., Co-chair

Doctoral Committee

Zhangsheng Yu, Ph.D., Co-chair

Hai Liu, Ph.D.

November 04, 2014

Yiqing Song, M.D., Sc.D.

DEDICATION

To My Family

ACKNOWLEDGMENTS

I would like to express sincere gratitude to my advisors Dr. Wanzhu Tu and Dr. Zhangsheng Yu for their constant guidance, encouragement and support in my Ph.D. study. I really appreciate the opportunity that they lead me to grow in this wonderful research area. Their guidance has not only trained my knowledge and expertise, but also cultivated me with open-mindedness, ability of critical thinking, and skills of effective communication which are essential for my future career. I have also learned from them the spirit of hard working, persistance, patience and creativity, which I will benefit from for the rest of my life.

I would like to thank the committee members for my thesis research, Dr. Hai Liu and Dr. Yiqing Song for their critical evaluations on my dissertation. I would also like to specially thank Dr. Changyu Shen and Dr. Xiaochun Li for their help and support during my dissertation work.

I feel quite grateful to this wonderful Biostatistics Ph.D. program, faculty and staff to provide the friendly and interdesciplinary research environment. I also must thank the classmates and friends who have helped and supported me. I would always remember the joy shared with them.

Finally, I sincerely thank my parents, for their unconditional love and support, as well as their kindness and optimism to make this journey come true.

Zangdong He

VARIABLE SELECTION AND STRUCTURAL DISCOVERY IN JOINT MODELS OF LONGITUDINAL AND SURVIVAL DATA

Joint models of longitudinal and survival outcomes have been used with increasing frequency in clinical investigations. Correct specification of fixed and random effects, as well as their functional forms is essential for practical data analysis. However, no existing methods have been developed to meet this need in a joint model setting. In this dissertation, I describe a penalized likelihood-based method with adaptive least absolute shrinkage and selection operator (ALASSO) penalty functions for model selection. By reparameterizing variance components through a Cholesky decomposition, I introduce a penalty function of group shrinkage; the penalized likelihood is approximated by Gaussian quadrature and optimized by an EM algorithm. The functional forms of the independent effects are determined through a procedure for structural discovery. Specifically, I first construct the model by penalized cubic B-spline and then decompose the B-spline to linear and nonlinear elements by spectral decomposition. The decomposition represents the model in a mixed-effects model format, and I then use the mixed-effects variable selection method to perform structural discovery. Simulation studies show excellent performance. A clinical application is described to illustrate the use of the proposed methods, and the analytical results demonstrate the usefulness of the methods.

Wanzhu Tu, Ph.D., Co-chair

Zhangsheng Yu, Ph.D., Co-chair

TABLE OF CONTENTS

LIST OF TABLES ix

LIST OF FIGURES xi

Chapter 1 Introduction 1

1.1 Joint Models of Longitudinal and Survival Outcomes 1

1.2 Simultaneous Variable Selection 4

1.3 Structural Discovery for Joint Models 6

1.4 Selection of Time-Varying Coefficients 7

Chapter 2 Selection of Fixed and Random Effects 10

2.1 Introduction 10

2.2 Method 12

2.2.1 Model Formulation 12

2.2.2 Variable Selection Using Penalized Likelihood 13

2.2.3 EM Algorithm for Optimization of the Penalized Likelihood . . . 16

2.2.4 Tuning Parameter Selection and Two-stage Estimation 20

2.3 Simulation Study 21

2.3.1 Data Generation 21

2.3.2 Simulation Results 25

2.4 Data Application 38

2.5 Discussion 42

Chapter 3 Structural Discovery 44

3.1 Introduction 44

3.2 Method 47

3.2.1 Model Formulation 47
3.2.2 Penalized Smoothing Splines 49
3.2.3 Structural Discovery Using Reparametrized Penalized Smoothing Splines 50
3.2.4 EM Algorithm for Optimization of the Penalized Likelihood . . . 53
3.2.5 Tuning Parameter Selection and Two-stage Estimation 57
3.3 Simulation Study 58
3.3.1 Data Generation 58
3.3.2 Simulation Results 59
3.4 Discussion 66
Chapter 4 Selection of Time-Varying Coefficients 68
4.1 Introduction 68
4.2 Method 70
4.2.1 Model Formulation 70
4.2.2 Representing the Model by Decomposed B-spline 71
4.2.3 Selection of Time-Varying Coefficients by Penalized Likelihood . 73
4.2.4 Optimization of the Penalized Likelihood 75
4.2.5 Tuning Parameter Selection and Two-stage Estimation 77
4.3 Simulation Study 78
4.3.1 Data Generation 78
4.3.2 Simulation results 79
4.4 Discussion 84
Chapter 5 Conclusion 86
BIBLIOGRAPHY 90
CURRICULUM VITAE

LIST OF TABLES

2.1 Selection frequency of mixed effects in longitudinal and survival components for Scenarios 1 to 4 28

2.2 Estimation of fixed effects $\beta_{1,j}$ in longitudinal component for Scenarios 1 to 4 29

2.3 Estimation of fixed effects $\beta_{2,j}$ in survival component for Scenarios 1 to 4 30

2.4 Estimation of random effects $\sqrt{D_{1kk}}$ and $\sqrt{D_{2kk}}$ in longitudinal and survival components for Scenarios 1 to 4. 31

2.5 Selection frequency of mixed effects in longitudinal and survival components for Scenario 5 32

2.6 Estimation of fixed effects $\beta_{1,j}$ and $\beta_{2,j}$ in longitudinal and survival components for Scenario 5 33

2.7 Estimation of random effects $\sqrt{D_{1kk}}$ and $\sqrt{D_{2kk}}$ in longitudinal and survival components for Scenario 5 34

2.8 Selection frequency of mixed effects in longitudinal and survival components for Scenario 6 35

2.9 Estimation of fixed effects $\beta_{1,j}$ and $\beta_{2,j}$ in longitudinal and survival components for Scenario 6 36

2.10 Estimation of random effects $\sqrt{D_{1kk}}$ and $\sqrt{D_{2kk}}$ in longitudinal and survival components for Scenario 6 37

2.11 Results for the heart failure patient data analysis. 40

3.1 Structural discovery accuracy in longitudinal and survival components for Scenarios 1 to 3 62

3.2 TAISE of longitudinal and survival components for Scenarios 1 to 3 . . 62

4.1 Selection frequency of time-invariant coefficients (TIC) 80

4.2 Selection frequency of time-varying coefficients (TVC) 81

4.3 TAISE in longitudinal and survival components for time-varying coefficients 81

4.4 Estimation results of nonzero time-invariant coefficients (TIC) 82

4.5 Estimation results of zero time-invariant coefficients (TIC) 83

LIST OF FIGURES

2.1 Residual plots for data application diagnostics. The circles are the standardized residuals. The black lines are the LOESS estimates. 41

3.1 Curve estimates in the longitudinal component for Scenario 1. 63

3.2 Curve estimates in the survival component for Scenario 1. 63

3.3 Curve estimates in the longitudinal component for Scenario 2. 64

3.4 Curve estimates in the survival component for Scenario 2. 64

3.5 Curve estimates in the longitudinal component for Scenario 3. 65

3.6 Curve estimates in the survival component for Scenario 3. 65

Chapter 1

Introduction

Statistical modeling has played an increasingly important role in modern scientific investigation. In biomedical research, a significant number of discoveries were made using innovative analytical models. But the validity of model-based scientific inquiry is usually contingent on the correct specification of the model. Failure to include relevant independent variables, for example, will result in questionable inference, while including irrelevant variables creates numerical instability and reduces analytical efficiency. Determination of the correct model structure based on observed data has therefore become an essential component of the modeling process. Ultimately, one hopes to achieve a parsimonious modeling structure without sacrificing predictive or explanatory power.

The objective of this dissertation is to develop a set of model selection tools for joint models of longitudinal and survival outcomes. In this chapter, I present my research questions, review the existing literature, and describe the general approach that I use in this research.

1.1 Joint Models of Longitudinal and Survival Outcomes

The concept of joint models was first proposed by Tsiatis and colleagues to characterize the longitudinal relationship between a disease marker and a time-to-event process (Wulfsohn and Tsiatis, 1997). Early applications of such models include HIV clinical trials that prospectively measure CD4 counts (or viral loads) and disease mortality (De Gruttola and Tu, 1994; Tsiatis et al., 1995). Here the repeatedly assessed CD4 counts are treated as longitudinal outcomes whereas HIV-related mortality is considered as the survival outcome,

with the purpose of better delineating the relationship between the two. A similar approach is used in studies of prostate cancer, where repeatedly measured prostate-specific antigen (PSA) levels are used as longitudinal outcomes and time to disease reoccurrence is used as the survival outcome (Wulfsohn and Tsiatis, 1997; Xu and Zeger, 2001a).

Joint models, in comparison with the traditional analysis of modeling one outcome at a time, represent a significantly improved analytical approach. Among other things, it affords an opportunity to investigate the intercorrelation, or mutual influences, of the longitudinal and survival outcomes. In aforementioned HIV example, counts of CD4 lymphocytes indicate the strength of host immunity against infectious pathogens, thus could be directly related to patient mortality, which in turn censors the CD4 measurement. Failure to accommodate the interdependency of the two outcomes thus not only deprives the possibility of exploring the between-outcome association, but also introduces additional biases in estimation in the presence of measurement errors and early study dropout, especially if the latter is caused by disease exacerbation as reflected by CD4 counts (Tsiatis and Davidian, 2004). To understand the limitations of the separate modeling approach, one only has to look at the traditional two-stage modeling process. In the first stage, a linear mixed-effects model is used to determine the mean levels of the longitudinal outcome; in the second stage, the predicted values from the longitudinal model in the first stage are fed into the survival model. Since deceased patients may have a different longitudinal outcome trajectory, compared to those who survived, the two-stage modeling approach could introduce a significant amount of bias into the estimation.

It is from this context that joint models are developed as an alternative modeling strategy. By simultaneously accommodating both outcomes, joint models create a structure that retains the natural correlations between the outcomes, thus alleviating the bias due to informative missing such as early study dropout. In practice, this implies improved prediction

accuracy of survival outcome based on longitudinal measurements. As noted in previous research, joint models generally have better efficiency in parameter estimation (Faucett and Thomas, 1996).

To link the two outcomes together, one often resorts to the use of a "shared latent process" (Wulfsohn and Tsiatis, 1997). For example, in the context of HIV study, the shared latent process is typically thought of as an unobserved disease progress that determines both host immunity (or disease severity), as indicated by CD4 counts, and risk of mortality. By depicting the latent process with a set of random effects and letting them be shared by longitudinal and survival models, one connects the two models and introduces a patient-specific measure of frailty. Such a joint model formulation has been successfully used in many clinical investigations and it is increasingly being recognized as a mainstay analytical method. Early application of joint model mainly focused on the HIV/AIDS trials (De Gruttola and Tu, 1994; Tsiatis et al., 1995), which remains an important tool in this area (Wu et al., 2010). Another major application of the joint model is in the area of oncology trials to evaluate the association between a patient's quality of life and time to event end point (Ibrahim et al., 2010), or in the cancer vaccine trials, to investigate the correlation between repeatedly measured immunologic outcomes and patients' survival outcomes, for example, patients' time to relapse (Brown and Ibrahim, 2003).

A practical barrier for a more widespread use of this effective analytical approach is the lack of tools for model construction. Specifically, there is no specific guidance on the inclusion and exclusion of independent effects (both fixed and random), the determination of the functional forms of which the independent variables take, and the inclusion of time interactions. In practice, these important questions are left to the analyst, who usually decides in an empirical fashion. Conflicting results may arise as a consequence.

Methodologically, there is no systematic study of model selection in a joint model setting. Determination of an appropriate model structure is by no means trivial, even in traditional model settings. The joint model structure has significantly magnified the challenge. The goal of this dissertation is to develop a new class of methods for constructing valid joint models. My research consists of three independent but interrelated topics: (1) Fixed and random effect selection; (2) Determination of functional form of an independent variable, also known as structural discovery; and (3) Selection of time-varying coefficients. The ultimate goal is to produce a set of data-driven tools to assist analysts construct joint models of longitudinal and survival outcomes.

1.2 Simultaneous Variable Selection

Variable selection has long been viewed as a necessary safeguard for model validity. In a joint model, variable selection has taken on an additional importance of justifying the simultaneous modeling formulation by testing the existence of the shared latent processes, as embodied by the random effects. As a result, variable selection for joint models typically includes the selection of both fixed and random effects.

Existing approaches for variable selection. There is a sizable literature on variable selection in generalized linear models and proportional hazard model settings. There are three general approaches for variable selection. First, a traditional method is to exhaustively compare all possible models based on a predefined criterion, typically an information-based criterion, such as the Akaike information criteria (AIC, Akaike 1974) or Bayesian information criterion (BIC, Schwarz 1978). This approach has been used widely in the last several decades and a number of statistical tests, such as the likelihood ratio test, wald-type test, or score test have been derived for variable selection, most notably in less complicated modeling settings. A clear limitation of this approach is that the heavy computational

burden of fitting of all candidate models. Perhaps for this reason, the method has never been extended to the joint modeling setting. The second approach is the stepwise variable selection method. Although it is computationally more efficient than the first approach, it does not search the entire model space, thus leaving open the possibility that the true model could be missed. When the stepwise approach is applied to mixed-effects models, the alternating procedure of fixing either the mean model for the fixed effects or the covariance structure for the random effects yields no unified tests for both types of effects. This could lead to erroneous results as it makes assumptions about the model by fixing part of its structure. An ideal variable selection method for selecting fixed and random effects is to simultaneously select the two parts and search though the full model space. The third approach is the penalized likelihood method. This is a data-driven method requiring less model assumptions and is computationally more efficient, and can perform simultaneous selection of fixed and random effects in a unified framework.

Penalized likelihood method. In this research, I take the third approach - penalized likelihood method - for variable selection in the joint modeling setting. Briefly, the penalized likelihood approach was proposed in the mid-1990's by Tibshirani (1996). He proposed a least absolute shrinkage and selection operator (LASSO) for fixed-effect variable selection. The "oracle" properties of the smoothly clipped absolute deviation (SCAD; Fan and Li, 2001) and the adaptive least absolute shrinkage and selection operator (ALASSO; Zou, 2006) further strengthened the applicability of this approach. The "oracle" property refers to the consistency between the selected model and the underlying true model. A number of variable selection methods based on the penalized likelihood approach have since been developed for the longitudinal and survival models, although separately. Most of the studies have focused on the selection of fixed effects. Even in the separate models, simultaneous selection of mixed effects presents a formidable challenge. It is not until 2010, simultaneous

selection of fixed and random effects in a linear mixed model setting has not been resolved (Bondell et al., 2010). More recently, Ibrahim et al. (2011) studied the mixed-effect variable selection in generalized linear mixed models. To the best of my knowledge, no work has been done for simultaneous selection of fixed and random effects in the joint model setting.

Variable selection in joint models. The first part of my dissertation concerns the development of a variable selection method to simultaneously select the mixed effects in a joint model setting. All things considered, this is not a trivial extension of the previous work, as the joint model structure is much more complicated than the separate model. The approach will clearly identify the connections between the two model components, and to simultaneously select the mixed effects in the two components of the joint model. I reparameterize the joint model to achieve the goal of selection by using a penalized likelihood method.

1.3 Structural Discovery for Joint Models

A logical question following variable selection is what functional form should a variable take. Traditionally, all variables enter the model in a linear form, despite the fact that in biological science, few factors have truly linear influences. I therefore ask how should one determine the functional form of an independent variable? Should an effect be linear, nonlinear, or partially linear? Linear effects are commonly assumed for convenience of model fitting and result interpretation. But modeling a nonlinear effect as linear is a form of model misspecification and may result in erroneous inference. On the other hand, specifying a linear effect using splines or other nonparametric techniques while the true effect is linear will result in reduced efficiency and difficulty in model interpretation.

A valid and efficient regression model requires correct specification of the effect pattern for each independent variable. If an independent variable has a linear effect, one would like to model it as such. Otherwise, if an independent variable's effect is nonlinear, one

wants to model it nonparametrically. A data-driven approach to the nonlinearity in an independent variable is often referred to as "structural discovery". Extensive studies have been done on estimating parameters in a pre-specified linear or nonlinear model, in the separate longitudinal or survival model settings. However, few studies have attempted to detect nonlinearity for the purpose of specifying the functional form of an independent variable. More recently, Zhang et al. (2011) proposed a method for structural discovery in a partially linear model setting.

To ensure the validity of statistical inference, structural discovery procedures are needed for joint models. Extensive literature search suggests that no work has been done in this front. Given its complicated model structure, an ideal simultaneous structural discovery tool should require minimal assumptions and must be implementable without exorbitant computing resources. The second part of my dissertation focuses on this task.

1.4 Selection of Time-Varying Coefficients

Time-varying coefficient models. A more recent extension of linear regression is the addition of time-varying coefficient, which depicts the effect of an independent variable on the outcome not as a constant but as a function of an independent variable (Hastie and Tibshirani, 1993). Such an extension has greatly enhanced the modeling flexibility and has been used to discover important, but nonlinear, biological influences that would have been missed by traditional analysis. For example, the effect of sodium-retaining hormone aldosterone on blood pressure may be dependent on the prevailing levels of extracellular fluid volume, as reflected by plasma renin activity. Varying coefficient model provides a flexible modeling framework to accommodate such interacting influences (Tu et al., 2014). But more often, the effects of certain independent variables on the outcome change over time, thus providing the incentive to model the effect as *a function of time*. Such a need

gives rise to time-varying coefficient models. Similarly, in survival analysis, one often has the need to model the time-dependent effect of an independent variable. For example, in an analysis of sexually transmitted infections, Yu et al. (2012) showed the effect of number of partners on infection acquisition tended to be age-dependent. In a childhood asthma study, the effect of airway reactivity measurement on the risk of wheezing also changed over time due to child growth (Yu et al., 2013).

Popular estimation methods for time-varying coefficients include kernel based local likelihood, smoothing spline and B-spline (Yan and Huang, 2012). Although the time-varying coefficient could uncover the temporal pattern of an independent variable effect, unnecessary nonparametric estimation makes it difficult to interpret the model and also lose some model efficiency. If the independent variable effect is constant over time, the model with time-invariant coefficients is more favorable for better interpretability and increased efficiency.

Selecting time-varying coefficient in joint models. In joint models, independent variables could interact with time, creating a need for time-varying coefficient, although the effects of the same independent variable on the two outcomes could take different functional forms. A statistical tool that helps to determine the functional forms would be very useful in such a modeling situation. Specifically, the tool should be able to consistently distinguish the independent variables with time-invariant or time-varying coefficients.

Model selection tools for time-varying coefficient model is in general limited. For longitudinal study, Wang et al. (2008) proposed a penalized likelihood method with SCAD penalty on the expanded nonparametric basis functions of coefficients. For the survival analysis, Yan and Huang (2012) proposed to use ALASSO to select time-invariant and time-varying coefficients, as well as excluding the zero coefficients in the Cox model. Careful literature search does not yield published work in selection of time-varying coefficient in joint model

settings. I therefore focus on the development of such a procedure in the third part of my dissertation.

Chapter 2

Selection of Fixed and Random Effects

2.1 Introduction

Longitudinal and survival data often arise together in clinical investigations. In a given subject, longitudinally measured clinical markers and patient survival are usually governed by the same latent disease process, and thus are correlated. Separate modeling for the longitudinal and survival outcomes could result in biases in parameter estimation (Faucett and Thomas, 1996). Joint models are therefore recommended to alleviate biases and to ensure valid inference concerning the correlation structure between the two outcomes. In the past two decades, joint models have been studied extensively: Wulfsohn and Tsiatis (1997) proposed a general framework in which the survival component was depicted by a proportional hazard model, and the longitudinal component was accommodated by a linear-growth-curve model. This basic structure was later extended by Xu and Zeger (2001b) to a variety of data situations. Other noteworthy method developments and significant data applications were presented by De Gruttola and Tu (1994), Nathoo and Dean (2008) and Albert and Shih (2010). Notably missing in this literature is variable selection. As in any modeling exercise, correct specification of the model and inclusion of the right independent variables are of essential importance, for the preservation of scientific validity. For joint models in particular, random variable selection serves the purpose of justifying the use of shared random effects connecting the longitudinal and survival components.

Traditionally, variable selection has been performed through model comparisons using information-based criteria, such as the Akaike and Bayesian information criteria (AIC and BIC). But such criteria are not always feasible in complex model settings where the number

of candidate models is large. As an alternative, penalized likelihood approach has gained popularity since the mid-1990's. Tibshirani (1996) proposed a least absolute shrinkage and selection operator (LASSO) for fixed-effect selection. Asymptotic "oracle" properties of the smoothly clipped absolute deviation (SCAD; Fan and Li, 2001) and the adaptive least absolute shrinkage and selection operator (ALASSO; Zou, 2006) have provided a theoretical assurance for mixed-effect selection. Along this line, Fan and Li (2004), Garcia et al. (2010) and Johnson et al. (2008) discussed the application of penalized likelihood method to select fixed effect variables in longitudinal model settings. Fan and Li (2002), Garcia et al. (2010) and Zhang and Lu (2007) discussed the selection of fixed effects in survival models. Extending these previous work, Bondell et al. (2010) proposed a method for selecting fixed and random effects in a linear mixed-effects model setting. Most recently, Ibrahim et al. (2011) studied the mixed-effects selection in generalized linear mixed models through an EM algorithm. To the best of our knowledge, no work has been done for simultaneous selection of fixed and random effects in a joint model setting with longitudinal and survival outcomes. To fill in this methodological gap, I propose a penalized likelihood method with ALASSO penalty for fixed and random effect selection in joint models. I optimize the penalized likelihood using an EM algorithm.

I illustrate the method by analyzing data from an observational study of heart failure patients. The study cohort included 1702 patients with diagnosed congestive heart failure (CHF) between Jan 1, 2004 and Dec 31, 2009, identified from a large electronic medical record system. The analytical objective is to assess the effects of medication adherence on disease exacerbation and on patient survival; I also like to assess the correlation between CHF exacerbation and patient mortality. Specifically, I considered two outcomes: the survival outcome is defined as the time from the first recorded CHF diagnosis to mortality, or to Dec 31, 2009, which ever comes first; the longitudinal outcomes are the repeatedly

measured B-type natriuretic peptide (BNP) levels. BNP is a commonly used bedside marker of CHF exacerbation; a higher BNP value indicates fluid volume overload in the left ventricle and increased mortality risk. (Morrison et al., 2002). Although the two outcomes can be modeled individually, separate modeling does not accommodate correlations between BNP and survival. In this research, I consider a joint modeling approach. I consider eight known risk factors and four interaction terms as candidate variables and develop an ALASSO procedure to select the independent variables. In particular, I consider random-effect selection as medical literature rarely avails information on the possible random slopes (e.g., the effect of an independent variable varies across subjects). Misspecification of fixed and random effects for the two outcome variables could result in erroneous inferences.

2.2 Method

2.2.1 Model Formulation

Suppose in a longitudinal study, I observe a survival outcome (t_i, δ_i), and repeated measurements of a continuous outcome $\boldsymbol{y}_i$, for subject $i = 1, \cdots, n$. Here t_i is the observed event time subject to right censoring, and δ_i is a failure indicator with $\delta_i = 1$ indicating the occurrence of an event of interest, and $\delta_i = 0$ indicating censoring, whereas $\boldsymbol{y}_i$ is an $n_i \times 1$ vector of the n_i repeated measurements. Let $\boldsymbol{X}_{1i} \in \mathbb{R}^{n_i \times p}$ and $\boldsymbol{Z}_{1i} \in \mathbb{R}^{n_i \times q}$ be the fixed and random covariate matrices for the longitudinal outcome, respectively. Similarly, I let $\boldsymbol{x}_{2i} \in \mathbb{R}_1^p$ and $\boldsymbol{z}_{2i} \in \mathbb{R}_1^q$ be the fixed and random covariate vectors for the survival outcome. Combining these observations I write $\boldsymbol{O_i} = (\boldsymbol{y}_i, \boldsymbol{X}_{1i}, \boldsymbol{Z}_{1i}, t_i, \delta_i, \boldsymbol{x}_{2i}, \boldsymbol{z}_{2i})$. I assume that the observations $\boldsymbol{O_i}$ are independent across subjects.

Without loss of generality, I herein consider a case where the longitudinal and survival components share the same set of fixed- and random-effect covariates. This model formula-

tion could easily be generalized to situations where the two components have different sets of covariates.

For the longitudinal outcome, I consider the following linear mixed-effects model:

$$\boldsymbol{y}_i = \boldsymbol{X}_{1i}\boldsymbol{\beta}_1 + \boldsymbol{Z}_{1i}\boldsymbol{\Gamma}_1\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \tag{2.1}$$

where $\boldsymbol{\beta}_1 = (\beta_{10}, \beta_{11}, \ldots, \beta_{1p})^T$ is the coefficient vector, and β_{10} is the intercept. $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \cdots, \varepsilon_{in_i})^T \sim N_{n_i}(0, \sigma^2 \boldsymbol{I}_{n_i})$ is the measurement error vector, and $\boldsymbol{b_i} \in \mathbb{R}_1^q$ is a $q-$dimensional random effect vector following a multivariate normal distribution $N_q(0, \boldsymbol{I}_q)$, with $\boldsymbol{I_q}$ as a $q \times q$ identity matrix. $\boldsymbol{\Gamma}_1$ is a $q \times q$ lower triangular matrix and $\boldsymbol{\Gamma}_1\boldsymbol{b}_i$ follows $N_q(0, \boldsymbol{D}_1)$. Thus $\boldsymbol{\Gamma}_1$ represents a Cholesky decomposition of the covariance matrix $\boldsymbol{D}_1$.

For the survival outcome, I consider a frailty model, defined as follows:

$$h(t_i) = h_0(t_i)\exp(\boldsymbol{x}_{2i}\boldsymbol{\beta}_2 + \boldsymbol{z}_{2i}\boldsymbol{\Gamma}_2\boldsymbol{b}_i), \tag{2.2}$$

where $h_0(t_i)$ is the baseline hazard function, and $\boldsymbol{\beta}_2 = (\beta_{21}, \ldots, \beta_{2p})^T$ is the coefficient vector. $\boldsymbol{\Gamma}_2\boldsymbol{b}_i$ follows $N_q(0, \boldsymbol{D}_2)$, and $\boldsymbol{\Gamma}_2$ is a Cholesky decomposition of the $q \times q$ matrix $\boldsymbol{D}_2$.

2.2.2 Variable Selection Using Penalized Likelihood

To select fixed and random effects, I propose a penalized likelihood to simultaneously identify the non-zero elements in $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\Gamma}_1\boldsymbol{b}_i, \boldsymbol{\Gamma}_2\boldsymbol{b}_i)$. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2, \boldsymbol{\phi})$ be the collection of all the unknown parameters, where $\boldsymbol{\phi}$ denotes parameters other than $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2)$. Writing the density function of $(\boldsymbol{y_i}, t_i, \boldsymbol{b_i})$ as $f(\boldsymbol{y}_i, t_i, \boldsymbol{b}_i | \boldsymbol{X}_{1i}, \boldsymbol{Z}_{1i}, \boldsymbol{x}_{2i}, \boldsymbol{z}_{2i}, h_0(t_i), \delta_i, \boldsymbol{\theta})$, I have the following log-marginal likelihood for $\boldsymbol{\theta}$:

$$l_o(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \int f_y(\boldsymbol{y}_i | \boldsymbol{X}_{1i}, \boldsymbol{Z}_{1i}, \boldsymbol{b}_i, \boldsymbol{\theta}) f_s(t_i, \delta_i | \boldsymbol{x}_{2i}, \boldsymbol{z}_{2i}, h_0(t_i), \boldsymbol{b}_i, \boldsymbol{\theta}) f_b(\boldsymbol{b}_i) d\boldsymbol{b}_i, \tag{2.3}$$

where $f_b(\boldsymbol{b_i})$ is a $q-$variate normal density function for $\boldsymbol{b_i}$. Functions $f_s(\cdot)$ and $f_y(\cdot)$ are the conditional density functions of the survival time and repeated measurements when $\boldsymbol{b_i}$ is given, respectively. I note that in the absence of restrictions on the baseline hazard $h_0(t_i)$, the maximum of the marginal likelihood is infinity. To remedy the deficiency, one could parameterize $h_0(t_i)$ with a parametric distribution. For example, a natural choice is to use a Weibull distribution with a baseline hazard $h_0(t_i) = \alpha\lambda t_i^{\alpha-1}$, where α is the shape parameter and λ is the scale parameter. Alternatively, one could use a piece-wise constant baseline hazard by dividing the study period into m intervals and assuming $h_0(t)$ to be a constant within each interval as $h_0(t) = h_k, t_{k-1} < t \leq t_k, k = 1 \ldots m$, where t_ks are knots defining the intervals. This piece-wise constant baseline hazard have been shown to perform well by Feng et al. (2005).

To select fixed and random effects simultaneously, I consider a penalized likelihood $PL(\boldsymbol{\theta}) = \frac{1}{n} l_o(\boldsymbol{\theta}) - \kappa_{\lambda_1}(\boldsymbol{\beta_1}) - \kappa_{\lambda_2}(\boldsymbol{\beta_2}) - \kappa_{\lambda_3}(\boldsymbol{D_1}) - \kappa_{\lambda_4}(\boldsymbol{D_2})$. The penalty terms $\kappa_{\lambda_1}(\boldsymbol{\beta_1})$ and $\kappa_{\lambda_2}(\boldsymbol{\beta_2})$ control for the sparsity of estimates of $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ so that the fixed effects are selected. The penalty terms $\kappa_{\lambda_3}(\boldsymbol{D_1})$ and $\kappa_{\lambda_4}(\boldsymbol{D_2})$ control for the sparsity of estimates of $\boldsymbol{D_1}$ and $\boldsymbol{D_2}$ to select the random effects. The penalty functions $\kappa_{\lambda_j}(\cdot)$, for $j = 1, 2, 3, 4$, could be the adaptive LASSO, or the smoothly clipped absolute deviation (SCAD). For the fixed-effect selection, I define the adaptive LASSO penalties as $\kappa_{\lambda_1}(\boldsymbol{\beta_1}) = \lambda_1 \sum_{j=1}^{p} \omega_{\beta_{1j}} |\beta_{1j}|$ and $\kappa_{\lambda_2}(\boldsymbol{\beta_2}) = \lambda_2 \sum_{k=1}^{p} \omega_{\beta_{2k}} |\beta_{2k}|$, where λ_1 and λ_2 are tuning parameters that control the degree of penalties; $\omega_{\beta_{1j}}, \omega_{\beta_{2k}}$ are the corresponding positive weights for penalties $|\beta_{1j}|$ and $|\beta_{2k}|$. The summation in $\kappa_{\lambda_1}(\boldsymbol{\beta_1}) = \lambda_1 \sum_{j=1}^{p} \omega_{\beta_{1j}} |\beta_{1j}|$ starts from 1 as I am not interested in selecting intercept β_{10}. Some of the estimates of $\hat{\beta}_{1j}$ and $\hat{\beta}_{2k}$ will be zero since $|\beta_{1k}|$ and $|\beta_{2k}|$ are singular when $|\beta_{1j}| = 0$ and $|\beta_{2k}| = 0$.

For the random-effect selection, I note that $\boldsymbol{D_1} = \boldsymbol{\Gamma_1 \Gamma_1^T}$ and $\boldsymbol{D_2} = \boldsymbol{\Gamma_2 \Gamma_2^T}$. Let $\boldsymbol{\gamma}_{1m}$ and $\boldsymbol{\gamma}_{2l}$ be the mth and lth row vectors of $\boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2$, respectively. In fact, $\boldsymbol{\gamma}_{1m} \boldsymbol{\gamma}_{1m}^T = \boldsymbol{D}_{1mm}$

and $\boldsymbol{\gamma}_{2l}\boldsymbol{\gamma}_{2l}^T = \boldsymbol{D}_{2ll}$ are the variance components of the mth and lth elements of the random effects $\boldsymbol{\Gamma_1 b_i}$ and $\boldsymbol{\Gamma_2 b_i}$. I form the penalty terms for the random effects in a group manner so that the estimates of elements of the entire vectors $\boldsymbol{\gamma}_{1m}$ and $\boldsymbol{\gamma}_{2l}$ are either all zero or at least one of the estimates is non-zero. The group penalties on $\boldsymbol{\gamma_{1m}}$ and $\boldsymbol{\gamma_{2l}}$ will ensure selection for the covariance structure due to the following connection of covariance matrices $\boldsymbol{D_1, D_2}$ and the Cholesky decomposition matrices $\boldsymbol{\Gamma_1, \Gamma_2}$ (Wang et al., 2010):

$$\begin{aligned} \boldsymbol{\gamma_{1m}} = 0 \Leftrightarrow D_{1mm} = 0, D_{1mh} = D_{1hm} = 0, \quad \forall h \\ \boldsymbol{\gamma_{2l}} = 0 \Leftrightarrow D_{2ll} = 0, D_{2lh} = D_{2hl} = 0, \quad \forall h. \end{aligned} \tag{2.4}$$

From (2.4), it follows that if $\boldsymbol{\gamma_{1m}} = 0$, then the diagonal element D_{1mm}, the variance of the random effect $(\boldsymbol{\Gamma_1 b_i})_m$, is zero. Furthermore, for any $h \neq m$, the off-diagonal element $D_{1mh} = D_{1hm} = 0$ implies that the covariance between $(\boldsymbol{\Gamma_1 b_i})_m$ and all other random effects are zero. Thus, the random effect $(\boldsymbol{\Gamma_1 b_i})_m$ in longitudinal component is to be excluded from the model and the positive-definiteness of $\boldsymbol{D_1}$ will be preserved. This applies to the random-effect selection in the survival component as well, which is to shrink the whole vector $\boldsymbol{\gamma_{2l}}$ to zero.

To perform group penalties on vectors $\boldsymbol{\gamma_{1m}}$ and $\boldsymbol{\gamma_{2m}}$, I first summarize the penalties using L_2-norm: $||\boldsymbol{\gamma}_{1m}|| = (\boldsymbol{\gamma}_{1m}\boldsymbol{\gamma}_{1m}^T)^{1/2}$ and $||\boldsymbol{\gamma}_{2l}|| = (\boldsymbol{\gamma}_{2l}\boldsymbol{\gamma}_{2l}^T)^{1/2}$ for $m, l = 2, \cdots, q$. Following Yuan and Lin (2006), the adaptive LASSO penalties are defined as: $\kappa_{\lambda_3}(\boldsymbol{D_1}) = \lambda_3 \sum_{m=2}^{q} \omega_{\boldsymbol{\gamma}_{1m}} ||\boldsymbol{\gamma}_{1m}||$ and $\kappa_{\lambda_4}(\boldsymbol{D_2}) = \lambda_4 \sum_{l=2}^{q} \omega_{\boldsymbol{\gamma}_{2l}} ||\boldsymbol{\gamma}_{2l}||$. I use adaptive LASSO penalties in the simulation study. Note that the summation starts from $m = 2, l = 2$, as I keep the random intercepts in both the longitudinal and survival components without eliminating the possible minimal within-cluster correlation. λ_3 and λ_4 are the positive tuning parameters, and $\omega_{\gamma_{1m}}, \omega_{\gamma_{2l}}$, are the positive weights associated with penalties on $||\boldsymbol{\gamma}_{1m}||$ and $||\boldsymbol{\gamma}_{2l}||$. Let $p(\boldsymbol{\theta}) = \lambda_1 \sum_{j=1}^{p} \omega_{\beta_{1j}} |\beta_{1j}| + \lambda_2 \sum_{k=1}^{p} \omega_{\beta_{2k}} |\beta_{2k}| + \lambda_3 \sum_{m=2}^{q} \omega_{\boldsymbol{\gamma_{1m}}} ||\boldsymbol{\gamma_{1m}}|| + \lambda_4 \sum_{m=2}^{q} \omega_{\boldsymbol{\gamma_{2l}}} ||\boldsymbol{\gamma_{2l}}||$,

and the penalized likelihood with the adaptive LASSO penalties can be written as

$$pl(\boldsymbol{\theta}) = \frac{1}{n} l_o(\boldsymbol{\theta}) - p(\boldsymbol{\theta}). \tag{2.5}$$

Penalized likelihood with SCAD penalties could be constructed by substituting the penalty terms in (2.5) using SCAD. The estimator of $\boldsymbol{\theta}$ can be obtained by maximizing (2.5).

2.2.3 EM Algorithm for Optimization of the Penalized Likelihood

To maximize the penalized likelihood (2.5), I use an EM algorithm. I start with the penalized log-complete likelihood for $(\boldsymbol{O_i}, \boldsymbol{b_i})$ for $i = 1, \cdots, n$, which is

$$\begin{aligned} pl_c(\boldsymbol{\theta}) =& \frac{1}{n} \sum_{i=1}^{n} \log f(\boldsymbol{y}_i, t_i, \delta_i, \boldsymbol{b}_i | \boldsymbol{\theta}) - p(\boldsymbol{\theta}) \\ =& \frac{1}{n} \sum_{i=1}^{n} \{\log[f_y(\boldsymbol{y}_i | \boldsymbol{b}_i, \boldsymbol{\theta})] + \delta_i \log[h(t_i | \boldsymbol{b}_i, \boldsymbol{\theta})] + \log[S(t_i | \boldsymbol{b}_i, \boldsymbol{\theta})] + \log[f_b(\boldsymbol{b}_i | \boldsymbol{\theta})]\} - p(\boldsymbol{\theta}). \end{aligned} \tag{2.6}$$

In Equation (2.6), $S(\cdot)$ is the survival function of t_i conditional on $\boldsymbol{b_i}$. Let $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)^T$ and $\boldsymbol{\omega} = (\omega_{\beta_{1j}}, \omega_{\beta_{2k}}, \omega_{\gamma_{1m}}, \omega_{\gamma_{2l}})^T$. I denote $\boldsymbol{g}_{c,i1} = (\boldsymbol{y}_i, \boldsymbol{X}_{1i}, \boldsymbol{Z}_{1i}, \boldsymbol{b}_i)$, $\boldsymbol{g}_{c,i2} = (t_i, \delta_i, \boldsymbol{x}_{2i}, \boldsymbol{z}_{2i}, \boldsymbol{b}_i)$ and $\boldsymbol{g}_{c,i} = (\boldsymbol{y}_i, \boldsymbol{X}_{1i}, \boldsymbol{Z}_{1i}, t_i, \delta_i, \boldsymbol{x}_{2i}, \boldsymbol{z}_{2i}, \boldsymbol{b}_i)$ as the complete data for longitudinal, survival and both components, respectively, and $\boldsymbol{g}_{o,i1} = (\boldsymbol{y}_i, \boldsymbol{X}_{1i}, \boldsymbol{Z}_{1i})$, $\boldsymbol{g}_{o,i2} = (t_i, \delta_i, \boldsymbol{x}_{2i}, \boldsymbol{z}_{2i})$ and $\boldsymbol{g}_{o,i} = (\boldsymbol{y}_i, \boldsymbol{X}_{1i}, \boldsymbol{Z}_{1i}, \boldsymbol{x}_{2i}, \boldsymbol{z}_{2i})$ as the corresponding observed data.

E-step

I first derive the E-step of the EM algorithm for the given $\boldsymbol{\lambda}$ and $\boldsymbol{\omega}$. Assuming that I have estimates $\boldsymbol{\theta}^{(s)}$ from the (s)th iteration of the maximization step, I take the expectation of the penalized log-complete likelihood conditional on $\boldsymbol{\theta}^{(s)}$ and $\boldsymbol{g}_{oi}$, for $i = 1, \ldots, n$ and obtain

the following penalized Q-function:

$$Q_{\lambda,\omega}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) = \frac{1}{n}\sum_{i=1}^{n}\{E[\log f_y(\boldsymbol{g}_{c,i1},\boldsymbol{\theta})|(\boldsymbol{g}_{o,i},\boldsymbol{\theta}^{(s)})] + E[\delta_i \log h(\boldsymbol{g}_{c,i2},\boldsymbol{\theta})|(\boldsymbol{g}_{o,i},\boldsymbol{\theta}^{(s)})] + E[\log S(\boldsymbol{g}_{c,i2},\boldsymbol{\theta})|(\boldsymbol{g}_{o,i},\boldsymbol{\theta}^{(s)})]\} - p(\boldsymbol{\theta}) + \frac{1}{n}\sum_{i=1}^{n} E[\log f_b(\boldsymbol{b}_i)|(\boldsymbol{g}_{o,i},\boldsymbol{\theta}^{(s)})]. \tag{2.7}$$

I write

$$E[H(\boldsymbol{b_i})|(\boldsymbol{g}_{o,i},\boldsymbol{\theta}^{(s)})] = \int H(\boldsymbol{b_i}) f_b(\boldsymbol{b}_i|\boldsymbol{g}_{o,i},\boldsymbol{\theta}^{(s)}) d\boldsymbol{b}_i, \tag{2.8}$$

for each of $H(\boldsymbol{b_i}) = \log f_y(\boldsymbol{g}_{c,i1},\boldsymbol{\theta})$, $H(\boldsymbol{b_i}) = \delta_i \log h(\boldsymbol{g}_{c,i2},\boldsymbol{\theta})$, and $H(\boldsymbol{b_i}) = \log S(\boldsymbol{g}_{c,i2},\boldsymbol{\theta})$. Because integral (2.8) is intractable, I approximate it by using a multivariate Gaussian quadrature method (Pinheiro and Bates, 1995). Since $\boldsymbol{b_i} \sim N(0, \boldsymbol{I}_q)$, if I choose k quadrature points in each dimension, there will be k^q vector nodes of $q \times 1$ dimension. Let $\boldsymbol{b}_l^{'} = (b_{l,1}^{'}, b_{l,2}^{'}, \cdots, b_{l,q}^{'})$ denote the lth node, and w_l denote the corresponding quadrature weight, for $l = 1, \cdots, k^q$, integral in (2.8) can be approximated by

$$\tilde{E}\{H(\boldsymbol{b_i})|(\boldsymbol{g}_{o,i},\boldsymbol{\theta}^{(s)})\} \approx \sum_{l=1}^{k^q} w_l H(\boldsymbol{b_l'}) f_b(\boldsymbol{b_l'}|\boldsymbol{g}_{o,i},\boldsymbol{\theta}^{(s)}). \tag{2.9}$$

I therefore obtain the approximated penalized Q-function in the $(s+1)$th iteration

$$\tilde{Q}_{\lambda,\omega}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) = \frac{1}{n}\sum_{i=1}^{n}\{\tilde{E}[\log f_y(\boldsymbol{g}_{c,i1},\boldsymbol{\theta})|(\boldsymbol{g}_{o,i},\boldsymbol{\theta}^{(s)})] + \tilde{E}[\delta_i \log h(\boldsymbol{g}_{c,i2},\boldsymbol{\theta})|(\boldsymbol{g}_{o,i},\boldsymbol{\theta}^{(s)})] + \tilde{E}[\log S(\boldsymbol{g}_{c,i2},\boldsymbol{\theta})|(\boldsymbol{g}_{o,i},\boldsymbol{\theta}^{(s)})]\} - p(\boldsymbol{\theta}). \tag{2.10}$$

The last term $\frac{1}{n}\sum_{i=1}^{n} E[\log f_b(\boldsymbol{b}_i)|(\boldsymbol{g}_{o,i},\boldsymbol{\theta}^{(s)})$ in (2.7) does not involve any unknown parameters, thus could be omitted from the optimization.

M-step

I maximize (2.10) with respect to the fixed- and random-effect parameters alternatively. When $(\boldsymbol{\Gamma_1, \Gamma_2, \phi})$ are fixed, I maximize (2.10) with respect to $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, and the penalty function involving L_1 penalty terms can be solved by applying the LARS/LASSO algorithm (Efron et al., 2004) and the SCAD penalties could be solved according to Fan and Li (2001). When $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\phi})$ are fixed, I maximize (2.10) with respect to $\boldsymbol{(\Gamma_1, \Gamma_2)}$. Following Lin and Zhang (2006) and Wang et al. (2010), I transform the optimization problem to a two-step equivalent objective function involving quadratic penalty term that is easier to solve. Specifically, let

$$\begin{aligned}\tilde{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) =& \frac{1}{n}\sum_{i=1}^{n}\{\tilde{E}[\log f_y(\boldsymbol{g}_{c,i1}, \boldsymbol{\theta})|(\boldsymbol{g}_{o,i}, \boldsymbol{\theta}^{(s)})] + \tilde{E}[\delta_i \log h(\boldsymbol{g}_{c,i2}, \boldsymbol{\theta})|(\boldsymbol{g}_{o,i}, \boldsymbol{\theta}^{(s)})] \\ &+ \tilde{E}[\log S(\boldsymbol{g}_{c,i2}, \boldsymbol{\theta})|(\boldsymbol{g}_{o,i}, \boldsymbol{\theta}^{(s)})]\},\end{aligned}$$

then for any given $\hat{\boldsymbol{\beta}}$ and $(\boldsymbol{\lambda}, \boldsymbol{\omega})$, the following two optimization problems with respect to $\boldsymbol{\gamma}$s achieve the same solution:

$$\tilde{Q}(\hat{\boldsymbol{\beta}}, \boldsymbol{\Gamma_1, \Gamma_2}|\boldsymbol{\theta}^{(s)}) - \lambda_3 \sum_{m=2}^{q} \omega_{\gamma_{1m}} ||\boldsymbol{\gamma_{1m}}|| - \lambda_4 \sum_{l=2}^{q} \omega_{\gamma_{2l}} ||\boldsymbol{\gamma_{2l}}|| \tag{2.11}$$

$$\tilde{Q}(\hat{\boldsymbol{\beta}}, \boldsymbol{\Gamma_1, \Gamma_2}|\boldsymbol{\theta}^{(s)}) - \sum_{m=2}^{q} \zeta_{1m}^2 - \frac{1}{4}\sum_{m=2}^{q} \frac{(\lambda_3 \omega_{\gamma_{1m}})^2}{\zeta_{1m}^2} ||\boldsymbol{\gamma_{1m}}||^2 - \sum_{l=2}^{q} \eta_{2l}^2 - \frac{1}{4}\sum_{l=2}^{q} \frac{(\lambda_4 \omega_{\gamma_{2l}})^2}{\eta_{2l}^2} ||\boldsymbol{\gamma_{2l}}||^2. \tag{2.12}$$

Let $(\hat{\boldsymbol{\gamma}}_{1m}, \hat{\boldsymbol{\gamma}}_{2l})$ be the maximizer of (2.11), and $(\tilde{\zeta}_{1m}, \tilde{\boldsymbol{\gamma}}_{1m}, \tilde{\eta}_{2l}, \tilde{\boldsymbol{\gamma}}_{2l})$ be the maximizer of (2.12), then I have

$$\hat{\boldsymbol{\gamma}}_{\boldsymbol{1m}} = \tilde{\boldsymbol{\gamma}}_{\boldsymbol{1m}}, \hat{\boldsymbol{\gamma}}_{\boldsymbol{2l}} = \tilde{\boldsymbol{\gamma}}_{\boldsymbol{2l}} \tag{2.13}$$

$$\tilde{\zeta}_{1m} = \sqrt{\frac{\lambda_3 \omega_{\gamma_{1m}}}{2} ||\hat{\boldsymbol{\gamma}}_{\boldsymbol{1m}}||}, \tilde{\eta}_{2l} = \sqrt{\frac{\lambda_4 \omega_{\gamma_{2l}}}{2} ||\hat{\boldsymbol{\gamma}}_{\boldsymbol{2l}}||}. \tag{2.14}$$

Equations (2.13) and (2.14) imply that one can optimize (2.12) iteratively with respect to $(\boldsymbol{\gamma_{1m}}, \boldsymbol{\gamma_{2l}})$ and (ζ_{1m}, η_{2l}), instead of directly maximizing (2.12). Maximizing (2.12) with respect to $(\boldsymbol{\gamma_{1m}}, \boldsymbol{\gamma_{2l}})$ when (ζ_{1m}, η_{2l}) is given is similar to a generalized ridge regression. When $(\boldsymbol{\gamma_{1m}}, \boldsymbol{\gamma_{2l}})$ is given, (ζ_{1m}, η_{2l}) could be easily computed from (2.14).

Let $\boldsymbol{\Theta} = (\boldsymbol{\theta}, \zeta_{1m}, \eta_{2l})$, where $\boldsymbol{\theta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2, \boldsymbol{\phi})$ are defined in section (2.2.2). I propose the expectation conditional maximization procedures to optimize the penalized likelihood as follows:

1. Initialize $(\boldsymbol{\beta_1^{(0)}}, \boldsymbol{\beta_2^{(0)}}, \boldsymbol{\gamma_{1m}^{(0)}}, \zeta_{1m}^{(0)}, \boldsymbol{\gamma_{2l}^{(0)}}, \eta_{2l}^{(0)}, \boldsymbol{\phi^{(0)}})$ with some plausible values.

2. For iteration s, update $\boldsymbol{\beta_1}, \boldsymbol{\beta_2}$ by adaptive LASSO,

$$\begin{aligned}\boldsymbol{\beta_1}^{(s)}, \boldsymbol{\beta_2}^{(s)} = \underset{\boldsymbol{\beta_1}, \boldsymbol{\beta_2}}{argmax}\, \tilde{Q}(\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \boldsymbol{\hat{\Gamma}_1^{(s-1)}}, \boldsymbol{\hat{\Gamma}_2^{(s-1)}}, \boldsymbol{\hat{\phi}^{(s-1)}} | \boldsymbol{\hat{\beta}_1^{(s-1)}}, \boldsymbol{\hat{\beta}_2^{(s-1)}}, \\ \boldsymbol{\hat{\Gamma}_1^{(s-1)}}, \boldsymbol{\hat{\Gamma}_2^{(s-1)}}, \boldsymbol{\hat{\phi}^{(s-1)}}) - \lambda_1 \sum_{j=1}^{p} \omega_{\beta_{1j}} |\beta_{1j}| - \lambda_2 \sum_{k=1}^{p} \omega_{\beta_{2k}} |\beta_{2k}|.\end{aligned}$$

3. Update $\boldsymbol{\gamma_{1m}}, \boldsymbol{\gamma_{2l}}$:

$$\begin{aligned}\boldsymbol{\gamma_{1m}}^{(s)}, \boldsymbol{\gamma_{2l}}^{(s)} = \underset{\gamma_{1m}, \gamma_{2l}}{argmax}\, \tilde{Q}(\boldsymbol{\hat{\beta}_1^{(s)}}, \boldsymbol{\hat{\beta}_2^{(s)}}, \boldsymbol{\Gamma_1}, \boldsymbol{\Gamma_2}, \boldsymbol{\hat{\phi}^{(s-1)}} | \boldsymbol{\hat{\beta}_1^{(s)}}, \boldsymbol{\hat{\beta}_2^{(s)}}, \boldsymbol{\hat{\Gamma}_1^{(s-1)}}, \boldsymbol{\hat{\Gamma}_2^{(s-1)}}, \boldsymbol{\hat{\phi}^{(s-1)}}) \\ - \frac{1}{4} \sum_{m=2}^{q} \frac{(\lambda_3 \omega_{\boldsymbol{\gamma_{1m}}})^2}{(\zeta_{1m}^{(s-1)})^2} ||\boldsymbol{\gamma_{1m}}||^2 - \frac{1}{4} \sum_{l=2}^{q} \frac{(\lambda_4 \omega_{\boldsymbol{\gamma_{2l}}})^2}{(\eta_{2l}^{(s-1)})^2} ||\boldsymbol{\gamma_{2l}}||^2.\end{aligned}$$

4. Update ζ_{1m}, η_{2l}:

$$\zeta_{1m}^{(s)} = \sqrt{\frac{\lambda_{\gamma_1} \omega_{\gamma_{1m}}}{2} ||\boldsymbol{\gamma_{1m}^{(s)}}||}, \eta_{2l}^{(s)} = \sqrt{\frac{\lambda_{\gamma_2} \omega_{\gamma_{2l}}}{2} ||\boldsymbol{\gamma_{2l}^{(s)}}||}.$$

5. Update $\boldsymbol{\phi}$:

$$\boldsymbol{\phi} = \underset{\boldsymbol{\phi}}{argmax}\, \tilde{Q}(\boldsymbol{\hat{\beta}_1^{(s)}}, \boldsymbol{\hat{\beta}_2^{(s)}}, \boldsymbol{\hat{\Gamma}_1^{(s)}}, \boldsymbol{\hat{\Gamma}_2^{(s)}}, \boldsymbol{\phi} | \boldsymbol{\hat{\beta}_1^{(s)}}, \boldsymbol{\hat{\beta}_2^{(s)}}, \boldsymbol{\hat{\Gamma}_1^{(s)}}, \boldsymbol{\hat{\Gamma}_2^{(s)}}, \boldsymbol{\hat{\phi}^{(s-1)}}).$$

6. Terminate the iteration when $max|\boldsymbol{\Theta}^{(s)} - \boldsymbol{\Theta}^{(s-1)}|$ are small enough. Otherwise, let $s = s + 1$ and go back to step 2.

Before updating parameters in each step, the corresponding $\tilde{Q}$ function is approximated by Gaussian quadrature in the E-step. To improve computation stability, smaller subset of $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2, \boldsymbol{\phi})$ could be updated iteratively. I could update $\boldsymbol{\beta}_1$ when $(\boldsymbol{\beta}_2, \boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2, \boldsymbol{\phi})$ is fixed, and then update $\boldsymbol{\beta}_2$ when $(\boldsymbol{\beta}_1, \boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2, \boldsymbol{\phi})$ is fixed, and sequentially for $\boldsymbol{\Gamma}_1$, $\boldsymbol{\Gamma}_2$, and $\boldsymbol{\phi}$ when other parameters are fixed. It is at the price of more iterations. The typical values for the weights are selected as: $\omega_{\beta_{1j}} = |\hat{\beta}^*_{1j}|^{-1}, \omega_{\beta_{2k}} = |\hat{\beta}^*_{2k}|^{-1}, \omega_{\gamma_{1m}} = \sqrt{m}||\hat{\boldsymbol{\gamma}}_{\mathbf{1}\boldsymbol{m}}{}^*||^{-1}, \omega_{\gamma_{2l}} = \sqrt{l}||\hat{\boldsymbol{\gamma}}_{\mathbf{2}\boldsymbol{l}}{}^*||^{-1}$, where $\hat{\beta}^*_{1j}, \hat{\beta}^*_{2k}, \hat{\boldsymbol{\gamma}}_{\mathbf{1}\boldsymbol{m}}{}^*, \hat{\boldsymbol{\gamma}}_{\mathbf{2}\boldsymbol{l}}{}^*$ are the unpenalized MLEs (Ibrahim et al., 2011; Zou, 2006) and $\sqrt{m}, \sqrt{l}$ are the normalizing constants for penalty parameters γ_{1m}, γ_{2l} to accommodate the varying sizes of $\boldsymbol{\gamma}_{\mathbf{1}\boldsymbol{m}}, \boldsymbol{\gamma}_{\mathbf{2}\boldsymbol{l}}$.

2.2.4 Tuning Parameter Selection and Two-stage Estimation

A data-driven method for determining tuning parameters is essential for variable selection. Criteria such as generalized cross-validation, k-fold cross validation, AIC, BIC, or GIC have been used as the objective scores to minimize over a preselected grid of tuning parameters. BIC is known to be consistent in the model selection (Pu and Niu, 2006; Shao, 1997). Wang et al. (2009) showed that selecting tuning parameters via BIC consistently yielded the true model in the linear model setting. Ibrahim et al. (2011) showed that selecting tuning parameters for mixed-effects selection via BIC-type IC_Q criterion also consistently yielded true models in generalized linear mixed models; their simulation study further showed that the approach worked well in finite sample situations. Thus, I propose to use the BIC-type criterion to determine the values of tuning parameters, where

$$BIC_{\boldsymbol{\lambda}} = -2l_o(\hat{\boldsymbol{\theta}}) + \log(n) \times df_{\boldsymbol{\lambda}}, \tag{2.15}$$

In (2.15), $\hat{\boldsymbol{\theta}}$ are the estimators obtained from penalized likelihood under the given $\boldsymbol{\lambda}$, and $l_o(\hat{\boldsymbol{\theta}})$ is the value of the observed likelihood $l_o(\boldsymbol{\theta})$ at the estimates $\hat{\boldsymbol{\theta}}$. The solution is chosen to minimize the $BIC_{\boldsymbol{\lambda}}$ criterion. In this BIC-type criterion, the total sample size n is used. I take d, the total number of non-zero estimates of $\hat{\theta}$ as the degree of freedom $df_{\boldsymbol{\lambda}}$. In the linear model, d is an unbiased estimator of $df_{\boldsymbol{\lambda}}$. Our simulation shows this criterion works well, as suggested by Pu and Niu (2006).

To reduce the estimation bias, I propose a two-stage process. In the first stage, I focus on variable selection and use the penalized likelihood method to select the model that minimizes the BIC value. In the second stage, I re-estimate parameters using selected variables without penalty for selection, to reduce the estimation bias.

2.3 Simulation Study

2.3.1 Data Generation

I conduct a simulation study to examine the performance of the proposed method. I generate data under six different scenarios.

For Scenarios 1 to 4, I generate the longitudinal outcome Y_{ij} from the following model:

$$\begin{aligned} Y_{ij} =& 1 + 1X_{1ij,1} + 0X_{1ij,2} + 3X_{1ij,3} + 0X_{1ij,4} + b_{li,0} \\ &+ b_{li,1}Z_{1ij,1} + b_{li,2}Z_{1ij,2} + b_{li,3}Z_{1ij,3} + b_{li,4}Z_{1ij,4} + \epsilon_{ij}, \end{aligned} \tag{2.16}$$

and the failure time from a Weibull distribution with the hazard function:

$$\begin{aligned} \lambda_i(t) = \lambda_0(t) \exp(&1x_{2i,1} + 0x_{2i,2} + 0x_{2i,3} + 1x_{2i,4} \\ &+ b_{si,0} + b_{si,1}z_{2i,1} + b_{si,2}z_{2i,2} + b_{si,3}z_{2i,3} + b_{si,4}z_{2i,4}), \end{aligned} \tag{2.17}$$

for $i = 1, \ldots, 250, j = 1, \ldots, 5$, where $\lambda_0(t) = \alpha\lambda t^{\alpha-1}$ with $\alpha = 2$, and $\lambda = \exp(1) = 2.718$.

Random effect vector $\boldsymbol{b_i}$ is independently generated from $N(0, \boldsymbol{I_5})$. $\boldsymbol{b_{li}} = (b_{li,0}, b_{li,1}, b_{li,2}, b_{li,3}, b_{li,4})$ is obtained from $\boldsymbol{b_{li}} = \boldsymbol{\Gamma_1 b_i}$ and $\boldsymbol{b_{si}} = (b_{si,0}, b_{si,1}, b_{si,2}, b_{si,3}, b_{si,4})$ is obtained from $\boldsymbol{b_{si}} = \boldsymbol{\Gamma_2 b_i}$, where $\boldsymbol{\Gamma_1} = \sigma_D \boldsymbol{R_1}$ and $\boldsymbol{\Gamma_2} = \sigma_D \boldsymbol{R_2}$, with lower triangular matrix

$$\boldsymbol{R_1} = \left\{ \begin{array}{ccccc} 1 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array} \right\}^{\frac{1}{2}}$$

and

$$\boldsymbol{R_2} = \left\{ \begin{array}{ccccc} 1 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \end{array} \right\}^{\frac{1}{2}}$$

Covariates $X_{1ij,1} = Z_{1ij,1}, X_{1ij,2} = Z_{1ij,2}, X_{1ij,4} = Z_{1ij,4}$ and $x_{2i,1} = z_{2i,1}, x_{2i,2} = z_{2i,2}, x_{2i,4} = z_{2i,4}$ are generated as independent $N(0,1)$ variables; $X_{1ij,3} = Z_{1ij,3}$ and $x_{2i,3} = z_{2i,3}$ are binary variables with equal probability taking value 0 or 1. The measurement error $\epsilon_{ij} \sim i.i.d. N(0,1)$. Censoring time is independently generated from an exponential distribution to achieve the desired censoring percentage.

In Scenario 1, I set σ_D to $\sqrt{0.5}$ and censoring percentage to 30%; in Scenario 2, I set σ_D to $\sqrt{1}$ and censoring percentage to 30%; in Scenario 3, I set σ_D to $\sqrt{0.5}$ and censoring percentage to 10%; in Scenario 4: I set σ_D to $\sqrt{1}$ and censoring percentage to 10%.

I additionally simulate data settings where there are higher proportions of censoring (Scenario 5) and larger numbers of random effects (Scenario 6). For scenario 5, I generate the longitudinal outcome Y_{ij} from the following model:

$$\begin{aligned} Y_{ij} =& 1 + 1.5X_{1ij,1} + 2X_{1ij,2} + 0X_{1ij,3} + 0X_{1ij,4} + b_{li,0} \\ &+ b_{li,1}Z_{1ij,1} + b_{li,2}Z_{1ij,2} + b_{li,3}Z_{1ij,3} + b_{li,4}Z_{1ij,4} + \epsilon_{ij}, \end{aligned}$$

and the failure time from a Weibull distribution with the hazard function:

$$\begin{aligned} \lambda_i(t) = \lambda_0(t) \exp(&1.5x_{2i,1} + 2x_{2i,2} + 0x_{2i,3} + 0x_{2i,4} \\ &+ b_{si,0} + b_{si,1}z_{2i,1} + b_{si,2}z_{2i,2} + b_{si,3}z_{2i,3} + b_{si,4}z_{2i,4}), \end{aligned}$$

for $i = 1, \ldots, 800, j = 1, \ldots, 5$, where $\lambda_0(t) = \alpha\lambda t^{\alpha-1}$ with $\alpha = 2$, and $\lambda = \exp(1) = 2.718$.

Random effect $\boldsymbol{b_i}$ is independently generated from $N(0, \boldsymbol{I_5})$. $\boldsymbol{b_{li}} = (b_{li,0}, b_{li,1}, b_{li,2}, b_{li,3}, b_{li,4})$ is obtained by $\boldsymbol{b_{li} = \Gamma_1 b_i}$ and $\boldsymbol{b_{si}} = (b_{si,0}, b_{si,1}, b_{si,2}, b_{si,3}, b_{si,4})$ is obtained by $\boldsymbol{b_{si} = \Gamma_2 b_i}$, where

$$\boldsymbol{\Gamma_1} = \boldsymbol{\Gamma_2} = \sigma_D \left\{ \begin{array}{ccccc} 1 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array} \right\}^{\frac{1}{2}}$$

and $\sigma_D = \sqrt{0.5}$. Covariates $X_{1ij,1} = Z_{1ij,1}, X_{1ij,2} = Z_{1ij,2}, X_{1ij,3} = Z_{1ij,3}, X_{1ij,4} = Z_{1ij,4}$ and $x_{2i,1} = z_{2i,1}$, $x_{2i,2} = z_{2i,2}, x_{2i,3} = z_{2i,3}, x_{2i,4} = z_{2i,4}$ are generated as independent $N(0, 1)$ variables; The measurement error $\epsilon_{ij} \sim i.i.d.N(0, 1)$. The censoring time is independently generated from an exponential distribution to achieve a 60% censoring percentage.

In Scenario 6, I generate the longitudinal outcome Y_{ij} from the following model:

$$\begin{aligned} Y_{ij} =& 1 + 1.5X_{1ij,1} + 2X_{1ij,2} + 2.5X_{1ij,3} + 0X_{1ij,4} + 0X_{1ij,5} + 0X_{1ij,6} + 0X_{1ij,7} + \\ & b_{li,0} + b_{li,1}Z_{1ij,1} + b_{li,2}Z_{1ij,2} + b_{li,3}Z_{1ij,3} + b_{li,4}Z_{1ij,4} + b_{li,5}Z_{1ij,5} + b_{li,6}Z_{1ij,6} + \\ & b_{li,7}Z_{1ij,7} + \epsilon_{ij}, \end{aligned}$$

and the failure time from a Weibull distribution with the hazard function:

$$\begin{aligned} \lambda_i(t) = \lambda_0(t) \exp(& 1.5x_{2i,1} + 2x_{2i,2} + 2.5x_{2i,3} + 0x_{2i,4} + 0x_{2i,5} + 0x_{2i,6} + 0x_{2i,7} + \\ & b_{si,0} + b_{si,1}z_{2i,1} + b_{si,2}z_{2i,2} + b_{si,3}z_{2i,3} + b_{si,4}z_{2i,4} + b_{si,5}z_{2i,5} + \\ & b_{si,6}z_{2i,6} + b_{si,7}z_{2i,7}), \end{aligned}$$

for $i = 1, \ldots, 250, j = 1, \ldots, 5$, where $\lambda_0(t) = \alpha\lambda t^{\alpha-1}$ with $\alpha = 2$, and $\lambda = \exp(1) = 2.718$.

Random effect $\boldsymbol{b_i}$ is independently generated from $N(0, \boldsymbol{I_8})$. $\boldsymbol{b_{li}} = (b_{li,0}, b_{li,1}, b_{li,2}, b_{li,3}, b_{li,4}\ b_{li,5}, b_{li,6}, b_{li,7})$ is obtained by $\boldsymbol{b_{li}} = \boldsymbol{\Gamma_1 b_i}$ and $\boldsymbol{b_{si}} = (b_{si,0}, b_{si,1}, b_{si,2}, b_{si,3}, b_{si,4}, b_{si,5}, b_{si,6}, b_{si,7})$ is obtained by $\boldsymbol{b_{si}} = \boldsymbol{\Gamma_2 b_i}$, where

$$\boldsymbol{\Gamma_1} = \boldsymbol{\Gamma_2} = \sigma_D \left\{ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}^{\frac{1}{2}} \right\}$$

and $\sigma_D = \sqrt{0.5}$. Covariates $X_{1ij,1} = Z_{1ij,1}, X_{1ij,2} = Z_{1ij,2}, X_{1ij,3} = Z_{1ij,3}, X_{1ij,4} = Z_{1ij,4}$, $X_{1ij,5} = Z_{1ij,5}$, $X_{1ij,6} = Z_{1ij,6}, X_{1ij,7} = Z_{1ij,7}$ and $x_{2i,1} = z_{2i,1}$, $x_{2i,2} = z_{2i,2}, x_{2i,3} = z_{2i,3}, x_{2i,4} = z_{2i,4}$, $x_{2i,5} = z_{2i,5}$, $x_{2i,6} = z_{2i,6}, x_{2i,7} = z_{2i,7}$ are generated as independent $N(0,1)$ variables; The measurement error $\epsilon_{ij} \sim i.i.d.N(0,1)$. The censoring time is independently generated from an exponential distribution to ahieve a 30% censoring percentage.

For each scenario, I generate 100 data sets and apply the proposed method to select the non-zero fixed or random effects in the first-stage model. After obtaining the selected variables, I fit the second-stage model including only the selected effects. The tuning parameters $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are determined by minimizing the BIC criterion, as defined in (2.15). The model without variable selection is also fitted for comparison.

2.3.2 Simulation Results

For Scenarios 1 to 4, I present the fixed- and random-effect selection results in Table 2.1, fixed-effect estimation results in Tables 2.2 and 2.3, and random-effect estimation results in Table 2.4. For fixed effects, the average correct selection rates are 100% for both non-zero and zero effects in longitudinal component, and 100% for non-zero and 98% for zero effects in survival component. The longitudinal fixed-effect estimates do not show severe biases in the first-stage estimation, and the biases are further reduced to less than 1% in the second-stage estimation. The survival fixed-effect estimates show 15% to 25% biases in the first-stage estimation, and the biases are reduced to below 4% in the second-stage estimation.

For random effects, the average rates of correct selection are 100% for non-zero and 94% for zero effects in longitudinal component, and approximately 96% for non-zero and 90% for zero effects in survival component. For non-zero random effects, the estimates in longitudinal component have biases ranging from 8% to 17% in the first-stage estimation,

and the biases are reduced to below 6% in the second-stage estimation. The survival non-zero random effect estimates show up to 42% biases in the first-stage estimation; in the second stage, the biases are reduced to less than 8%. For zero random effects, both the first- and second-stage estimates in longitudinal component have biases below 2%. The survival zero random effect estimates generally have less than 10% biases in both stage estimations.

Simulation results for settings with higher proportions of censoring and larger numbers of random effects are reported in Tables 2.5, 2.6, 2.7, 2.8, 2.9 and 2.10. Briefly, I find that the probabilities of correct selection remain excellent for these data settings.

One consequence of including more random effects is the increased computing time. The complexity of Gaussian quadrature increases exponentially with the dimension of the random effect vector. In this research, I used 3 quadrature points. With 3 quadrature points, each data set in Scenarios 1 to 4 took approximately 20 minutes to complete the first stage variable selection under one tuning parameter; and it took another 10 minutes in the second stage estimation. When I increased the number of random effects to 8 (as in Scenario 6), the computation time increased to 10 and 5 hours, respectively. The computing time is estimated on a single CPU (Intel(R) Xeon(R) CPU E7- 4830 @ 2.13GHz) and 4 GB memory in the Unix system. The total computing time depended on the number of tuning parameters. Other factors, such as the shape of the likelihood function could also influence the approximation accuracy of Gaussian quadrature and the computing time.

Generally, mis-selection rate increases as the censoring rate increases or the variance magnitude σ_D decreases, since smaller variance σ_D means less resolution between non-zero and zero random effects. The mis-selection subsequently leads to larger estimation bias. The influence of censoring rate on selection accuracy is greater than that of variance. Increased number of random effects does not necessarily lead to worse selection accuracy, but it

tends to slightly increase estimation bias, which may be due to the reduced approximation accuracy of Gaussian quadrature method. The estimates from the model without variable selection generally have more biases than the second-stage estimates, especially for the zero effects. In summary, I contend that the proposed variable selection and estimation method works well even under high proportions of censoring and large number of random effects. The two stage procedure ensures good selection performance in the first stage and reduced biased parameter estimation in the second stage.

Table 2.1: Selection frequency of mixed effects in longitudinal and survival components for Scenarios 1 to 4

Fixed effect selection								
	Sel. Freq.(%) for Longitudinal component				Sel. Freq.(%) for Survival component			
Scenarios	$X_{1,1}$ Non-Zero	$X_{1,2}$ Zero	$X_{1,3}$ Non-Zero	$X_{1,4}$ Zero	$X_{2,1}$ Non-Zero	$X_{2,2}$ Zero	$X_{2,3}$ Zero	$X_{2,4}$ Non-Zero
1	100	0	100	0	100	1	3	100
2	100	0	100	0	100	3	4	100
3	100	0	100	0	100	0	1	100
4	100	0	100	0	100	1	1	100

Random effect selection								
	Sel. Freq.(%) for Longitudinal component				Sel. Freq.(%) for Survival component			
Scenarios	$Z_{1,1}$ Non-Zero	$Z_{1,2}$ Zero	$Z_{1,3}$ Non-Zero	$Z_{1,4}$ Zero	$Z_{2,1}$ Non-Zero	$Z_{2,2}$ Zero	$Z_{2,3}$ Zero	$Z_{2,4}$ Non-Zero
1	100	8	100	11	99	7	19	84
2	100	2	100	6	100	10	13	92
3	100	2	100	10	100	5	12	97
4	100	1	100	10	100	6	8	99

Table 2.2: Estimation of fixed effects $\beta_{1,j}$ in longitudinal component for Scenarios 1 to 4

Scenarios		$\hat{\beta}_{1,j} \pm SE$(Coverage probability) for Longitudinal component[a]				
		Intercept	$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	$X_{1,4}$
	True value β	1	1	0	3	0
1	W/O selection $\hat{\beta}$	1.002±0.076(87%)	1.007±0.075(81%)	-0.003	3.009±0.077(96%)	0.000
	1^{st} stage $\hat{\beta}$	1.006±0.071(94%)	0.989±0.072(88%)	0.000	2.988±0.074(97%)	0.000
	2^{nd} stage $\hat{\beta}$	0.999±0.068(95%)	1.003±0.067(85%)	0.000	3.003±0.070(98%)	0.000
2	W/O selection $\hat{\beta}$	0.995±0.108(76%)	0.998±0.111(66%)	-0.005	2.999±0.111(89%)	0.001
	1^{st} stage $\hat{\beta}$	1.002±0.099(79%)	0.979±0.103(74%)	0.000	2.979±0.096(89%)	0.000
	2^{nd} stage $\hat{\beta}$	0.994±0.108(82%)	0.996±0.108(75%)	0.000	2.997±0.094(94%)	0.000
3	W/O selection $\hat{\beta}$	1.000±0.076(84%)	1.005±0.075(80%)	-0.003	3.007±0.078(96%)	-0.001
	1^{st} stage $\hat{\beta}$	1.007±0.070(96%)	0.989±0.072(91%)	0.000	2.988±0.074(98%)	0.000
	2^{nd} stage $\hat{\beta}$	0.998±0.070(93%)	1.002±0.066(88%)	0.000	3.003±0.07(98%)	0.000
4	W/O selection $\hat{\beta}$	0.994±0.113(72%)	1.002±0.111(63%)	-0.005	3.003±0.109(89%)	0.001
	1^{st} stage $\hat{\beta}$	1.004±0.102(85%)	0.983±0.103(78%)	0.000	2.982±0.098(91%)	0.000
	2^{nd} stage $\hat{\beta}$	0.995±0.108(81%)	1.000±0.104(73%)	0.000	2.999±0.091(95%)	0.000

[a] $\hat{\beta}s$ are the averages of estimates over the 100 data sets; SE is the empirical standard error of the 100 $\hat{\beta}s$; For each data set, the 95% confidence interval based on the parameter and standard error estimates is calculated and the corresponding coverage probabilities for the true value over the 100 data sets are included in the parentheses. SE and coverage probability are only reported for non-zero variables.

Table 2.3: Estimation of fixed effects $\beta_{2,j}$ in survival component for Scenarios 1 to 4

Scenarios		$\hat{\beta}_{2,j} \pm SE$(Coverage probability) for Survival component[a]				
		Intercept	$X_{2,1}$	$X_{2,2}$	$X_{2,3}$	$X_{2,4}$
	True value β	-	1	0	0	1
1	W/O selection $\hat{\beta}$		1.140±0.275(86%)	-0.009	0.010	1.141±0.308(81%)
	1^{st} stage $\hat{\beta}$		0.854±0.224(59%)	-0.001	-0.005	0.853±0.255(62%)
	2^{nd} stage $\hat{\beta}$		1.000±0.263(89%)	0.001	0.014	0.996±0.285(80%)
2	W/O selection $\hat{\beta}$		1.110±0.370(79%)	-0.023	-0.013	1.137±0.424(79%)
	1^{st} stage $\hat{\beta}$		0.751±0.312(35%)	0.001	-0.007	0.778±0.348(40%)
	2^{nd} stage $\hat{\beta}$		0.990±0.417(83%)	0.017	0.015	1.030±0.636(81%)
3	W/O selection $\hat{\beta}$		1.157±0.256(84%)	0.005	0.040	1.145±0.250(88%)
	1^{st} stage $\hat{\beta}$		0.878±0.187(63%)	-0.001	-0.003	0.876±0.187(66%)
	2^{nd} stage $\hat{\beta}$		1.023±0.158(98%)	0.000	0.003	1.021±0.161(92%)
4	W/O selection $\hat{\beta}$		1.113±0.259(84%)	0.000	0.054	1.132±0.299(85%)
	1^{st} stage $\hat{\beta}$		0.792±0.207(53%)	0.000	0.001	0.810±0.245(51%)
	2^{nd} stage $\hat{\beta}$		1.008±0.187(91%)	0.003	0.002	1.013±0.239(86%)

[a] $\hat{\beta}s$ are the averages of estimates over the 100 data sets; SE is the empirical standard error of the 100 $\hat{\beta}s$; For each data set, the 95% confidence interval based on the parameter and standard error estimates is calculated and the corresponding coverage probabilities for the true value over the 100 data sets are included in the parentheses. SE and coverage probability are only reported for non-zero variables.

Table 2.4: Estimation of random effects $\sqrt{D_{1kk}}$ and $\sqrt{D_{2kk}}$ in longitudinal and survival components for Scenarios 1 to 4.

		$\sqrt{\hat{D}_{1kk}}$ for Longitudinal component[a]					$\sqrt{\hat{D}_{2kk}}$ for Survival component[a]				
Scenarios		$Intercept_1$	$Z_{1,1}$	$Z_{1,2}$	$Z_{1,3}$	$Z_{1,4}$	$Intercept_2$	$Z_{2,1}$	$Z_{2,2}$	$Z_{2,3}$	$Z_{2,4}$
	True value $\sqrt{D_{kk}}$	0.707	0.707	0	0.707	0	0.707	0.707	0	0	0.707
1	W/O selection $\sqrt{\hat{D}_{kk}}$	0.658	0.668	0.112	0.788	0.133	0.770	0.799	0.441	0.775	0.944
	1^{st} stage $\sqrt{\hat{D}_{kk}}$	0.824	0.763	0.003	0.771	0.001	0.710	0.556	0.031	0.073	0.491
	2^{nd} stage $\sqrt{\hat{D}_{kk}}$	0.695	0.690	0.017	0.735	0.020	0.698	0.718	0.054	0.170	0.708
3	W/O selection $\sqrt{\hat{D}_{kk}}$	0.658	0.665	0.102	0.790	0.130	0.757	0.793	0.342	0.644	0.886
	1^{st} stage $\sqrt{\hat{D}_{kk}}$	0.828	0.750	0.002	0.741	0.001	0.727	0.585	0.006	0.028	0.440
	2^{nd} stage $\sqrt{\hat{D}_{kk}}$	0.695	0.690	0.005	0.734	0.016	0.690	0.731	0.021	0.071	0.725
		$Intercept_1$	$Z_{1,1}$	$Z_{1,2}$	$Z_{1,3}$	$Z_{1,4}$	$Intercept_2$	$Z_{2,1}$	$Z_{2,2}$	$Z_{2,3}$	$Z_{2,4}$
	True value $\sqrt{D_{kk}}$	1	1	0	1	0	1	1	0	0	1
2	W/O selection $\sqrt{\hat{D}_{kk}}$	0.877	0.913	0.105	1.103	0.136	1.034	1.094	0.552	0.870	1.262
	1^{st} stage $\sqrt{\hat{D}_{kk}}$	1.143	1.085	0.000	1.140	0.000	0.923	0.715	0.043	0.017	0.580
	2^{nd} stage $\sqrt{\hat{D}_{kk}}$	0.941	0.955	0.004	1.035	0.010	0.970	0.995	0.083	0.146	1.059
4	W/O selection $\sqrt{\hat{D}_{kk}}$	0.873	0.909	0.097	1.093	0.135	1.000	1.046	0.391	0.714	1.163
	1^{st} stage $\sqrt{\hat{D}_{kk}}$	1.140	1.080	0.000	1.132	0.000	0.956	0.783	0.014	0.010	0.591
	2^{nd} stage $\sqrt{\hat{D}_{kk}}$	0.938	0.953	0.003	1.032	0.015	0.923	0.975	0.038	0.053	1.006

[a] $\sqrt{\hat{D}_{1kk}}$ and $\sqrt{\hat{D}_{2kk}}$ are the averages of estimates over the 100 data sets.

Table 2.5: Selection frequency of mixed effects in longitudinal and survival components for Scenario 5

Fixed effect selection							
Sel. Freq.(%) for Longitudinal component				Sel. Freq.(%) for Survival component			
$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	$X_{1,4}$	$X_{2,1}$	$X_{2,2}$	$X_{2,3}$	$X_{2,4}$
Non-Zero	Non-Zero	Zero	Zero	Non-Zero	Non-Zero	Zero	Zero
100	100	0	0	100	100	0	0

Random effect selection							
Sel. Freq.(%) for Longitudinal component				Sel. Freq.(%) for Survival component			
$Z_{1,1}$	$Z_{1,2}$	$Z_{1,3}$	$Z_{1,4}$	$Z_{2,1}$	$Z_{2,2}$	$Z_{2,3}$	$Z_{2,4}$
Non-Zero	Non-Zero	Zero	Zero	Non-Zero	Non-Zero	Zero	Zero
100	100	0	0	99	99	1	0

Table 2.6: Estimation of fixed effects $\beta_{1,j}$ and $\beta_{2,j}$ in longitudinal and survival components for Scenario 5

	$\hat{\beta}_{1,j} \pm SE$(Coverage probability) for Longitudinal component[a]				
	Intercept	$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	$X_{1,4}$
True value β	1	1.5	2	0	0
W/O selection $\hat{\beta}$	0.995±0.033(94%)	1.500±0.036(95%)	1.998±0.039(94%)	0.002	0.001
1^{st} stage $\hat{\beta}$	0.993±0.033(92%)	1.487±0.036(90%)	1.986±0.039(87%)	0.000	0.000
2^{nd} stage $\hat{\beta}$	0.999±0.029(95%)	1.505±0.034(95%)	2.001±0.036(91%)	0.000	0.000

	$\hat{\beta}_{2,j} \pm SE$(Coverage probability) for Survival component[a]				
	Intercept	$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	$X_{1,4}$
True value β	-	1.5	2	0	0
W/O selection $\hat{\beta}$		1.355±0.126(74%)	1.844±0.149(75%)	0.008	0.019
1^{st} stage $\hat{\beta}$		0.989±0.125(0%)	1.381±0.145(1%)	0.000	0.000
2^{nd} stage $\hat{\beta}$		1.348±0.130(67%)	1.823±0.152(73%)	0.000	0.000

[a] $\hat{\beta}s$ are the averages of estimates over the 100 data sets; SE is the empirical standard error of the 100 $\hat{\beta}s$; For each data set, the 95% confidence interval based on the parameter and standard error estimates is calculated and the corresponding coverage probabilities for the true value over the 100 data sets are included in the parentheses. SE and coverage probability are only reported for non-zero variables.

Table 2.7: Estimation of random effects $\sqrt{D_{1kk}}$ and $\sqrt{D_{2kk}}$ in longitudinal and survival components for Scenario 5

	$\sqrt{\hat{D}_{1kk}}$ for Longitudinal component[a]					$\sqrt{\hat{D}_{2kk}}$ for Survival component[a]				
	$Intercept_1$	$Z_{1,1}$	$Z_{1,2}$	$Z_{1,3}$	$Z_{1,4}$	$Intercept_2$	$Z_{2,1}$	$Z_{2,2}$	$Z_{2,3}$	$Z_{2,4}$
True value $\sqrt{D_{kk}}$	0.707	0.707	0.707	0	0	0.707	0.707	0.707	0	0
W/O selection $\sqrt{\hat{D}_{kk}}$	0.791	0.817	0.817	0.052	0.050	0.776	0.820	0.825	0.205	0.202
1^{st} stage $\sqrt{\hat{D}_{kk}}$	0.787	0.773	0.763	0.000	0.000	0.407	0.368	0.319	0.000	0.000
2^{nd} stage $\sqrt{\hat{D}_{kk}}$	0.682	0.692	0.696	0.000	0.000	0.638	0.665	0.674	0.004	0.000

[a] $\sqrt{\hat{D}_{1kk}}$ and $\sqrt{\hat{D}_{2kk}}$ are the averages of estimates over the 100 data sets.

Table 2.8: Selection frequency of mixed effects in longitudinal and survival components for Scenario 6

Fixed effect selection													
Sel. Freq.(%) for Longitudinal component							Sel. Freq.(%) for Survival component						
$X_{1,1}$ Non-Zero	$X_{1,2}$ Non-Zero	$X_{1,3}$ Non-Zero	$X_{1,4}$ Zero	$X_{1,5}$ Zero	$X_{1,6}$ Zero	$X_{1,7}$ Zero	$X_{2,1}$ Non-Zero	$X_{2,2}$ Non-Zero	$X_{2,3}$ Non-Zero	$X_{2,4}$ Zero	$X_{2,5}$ Zero	$X_{2,6}$ Zero	$X_{2,7}$ Zero
100	100	100	0	0	0	0	100	100	100	0	0	0	0

Random effect selection													
Sel. Freq.(%) for Longitudinal component							Sel. Freq.(%) for Survival component						
$Z_{1,1}$ Non-Zero	$Z_{1,2}$ Non-Zero	$Z_{1,3}$ Non-Zero	$Z_{1,4}$ Zero	$Z_{1,5}$ Zero	$Z_{1,6}$ Zero	$Z_{1,7}$ Zero	$Z_{2,1}$ Non-Zero	$Z_{2,2}$ Non-Zero	$Z_{2,3}$ Non-Zero	$Z_{2,4}$ Zero	$Z_{2,5}$ Zero	$Z_{2,6}$ Zero	$Z_{2,7}$ Zero
100	100	100	0	0	0	0	97	93	94	6	4	1	9

Table 2.9: Estimation of fixed effects $\beta_{1,j}$ and $\beta_{2,j}$ in longitudinal and survival components for Scenario 6

	$\hat{\beta}_{1,j} \pm SE$(Coverage probability) for Longitudinal component[a]							
	Intercept	$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	$X_{1,4}$	$X_{1,5}$	$X_{1,6}$	$X_{1,7}$
True value β	1	1.5	2	2.5	0	0	0	0
W/O selection $\hat{\beta}$	0.994±0.068(85%)	1.498±0.081(75%)	1.999±0.072(79%)	2.496±0.072(81%)	0.001	-0.004	0.000	-0.003
1^{st} stage $\hat{\beta}$	0.987±0.068(89%)	1.454±0.079(82%)	1.960±0.072(87%)	2.462±0.072(87%)	0.000	0.000	0.000	0.000
2^{nd} stage $\hat{\beta}$	0.994±0.064(87%)	1.497±0.076(82%)	1.995±0.074(85%)	2.496±0.073(86%)	0.000	0.000	0.000	0.000
	$\hat{\beta}_{2,j} \pm SE$(Coverage probability) for Survival component[a]							
		$X_{2,1}$	$X_{2,2}$	$X_{2,3}$	$X_{2,4}$	$X_{2,5}$	$X_{2,6}$	$X_{2,7}$
True value β		1.5	2	2.5	0	0	0	0
W/O selection $\hat{\beta}$		1.966±0.286(63%)	2.667±0.377(49%)	3.313±0.429(49%)	0.014	-0.025	0.011	0.035
1^{st} stage $\hat{\beta}$		1.039±0.249(20%)	1.495±0.331(28%)	1.897±0.370(30%)	0.000	0.000	0.000	0.000
2^{nd} stage $\hat{\beta}$		1.549±0.358(86%)	2.112±0.593(82%)	2.625±0.712(84%)	0.000	0.000	0.000	0.000

[a] $\hat{\beta}s$ are the averages of estimates over the 100 data sets; SE is the empirical standard error of the 100 $\hat{\beta}s$; For each data set, the 95% confidence interval based on the parameter and standard error estimates is calculated and the corresponding coverage probabilities for the true value over the 100 data sets are included in the parentheses. SE and coverage probability are only reported for non-zero variables.

Table 2.10: Estimation of random effects $\sqrt{D_{1kk}}$ and $\sqrt{D_{2kk}}$ in longitudinal and survival components for Scenario 6

	$\sqrt{\hat{D}_{1kk}}$ for Longitudinal component[a]							
	Intercept	$Z_{1,1}$	$Z_{1,2}$	$Z_{1,3}$	$Z_{1,4}$	$Z_{1,5}$	$Z_{1,6}$	$Z_{1,7}$
True value $\sqrt{D_{kk}}$	0.707	0.707	0.707	0.707	0	0	0	0
W/O selection $\sqrt{\hat{D}_{kk}}$	0.785	0.821	0.831	0.829	0.165	0.174	0.159	0.158
1^{st} stage $\sqrt{\hat{D}_{kk}}$	0.768	0.677	0.657	0.633	0.000	0.000	0.000	0.000
2^{nd} stage $\sqrt{\hat{D}_{kk}}$	0.628	0.669	0.693	0.699	0.000	0.000	0.000	0.000
	$\sqrt{\hat{D}_{2kk}}$ for Survival component[a]							
	Intercept	$Z_{2,1}$	$Z_{2,2}$	$Z_{2,3}$	$Z_{2,4}$	$Z_{2,5}$	$Z_{2,6}$	$Z_{2,7}$
True value $\sqrt{D_{kk}}$	0.707	0.707	0.707	0.707	0	0	0	0
W/O selection $\sqrt{\hat{D}_{kk}}$	1.037	1.074	1.155	1.167	0.574	0.552	0.535	0.685
1^{st} stage $\sqrt{\hat{D}_{kk}}$	0.511	0.431	0.412	0.382	0.002	0.004	0.000	0.025
2^{nd} stage $\sqrt{\hat{D}_{kk}}$	0.652	0.698	0.725	0.782	0.047	0.018	0.005	0.091

[a] $\sqrt{\hat{D}_{1kk}}$ and $\sqrt{\hat{D}_{2kk}}$ are the averages of estimates over the 100 data sets.

2.4 Data Application

To illustrate the method, I analyzed observational data from the CHF study. As previously stated, the main purpose of the investigation is to assess the effects of medication adherence on disease exacerbation and patient survival. For the survival outcome, I modeled the time from the first recorded CHF diagnosis to patient mortality, which could be censored on Dec 31, 2009. For the longitudinal outcome, I modeled the repeatedly measured BNP levels as markers of disease exacerbation. Because the distribution of BNP skewed strongly to the right, I used the logarithmic-transformed BNP (log(BNP)) in the model. Medication adherence, the independent variable of primary interest, was the average proportion of days covered (PDC) by all prescribed medications within each patient (Choudhry et al., 2009). Besides PDC, seven other risk factors were considered, including systolic blood pressure (SBP), diastolic blood pressure (DBP), BMI, gender, age at CHF diagnosis date (IndexAge), number of comorbidities (NumComorbid) and number of medications taken (NumMed). I also considered interactions among SBP, DBP, BMI, PDC and gender.

In the study sample, 58.3% of the subjects were females and the average BMI was 32.7 (kg/m^2). The average age for the study cohort at the CHF diagnosis date was 62.7 years. On average, the study subjects had 5.1 comorbidities and took 8.4 medications with a mean PDC of 0.327. Among the covariates, concurrently measured SBP (mean: 134.8mmHg; SD: 24.2 mmHg) and DBP (mean: 77.0mmHg; SD: 16.0 mmHg) were recorded at the time of BNP assessment; the remaining variables were collected as baseline covariates. The censoring percentage was 64.1%, and median time to death was 4115 days (11.3 years).

For longitudinally measured BNP levels, I use linear mixed-effects model $\log(BNP)_{ij} = \boldsymbol{x}_{1,ij}\boldsymbol{\beta}_1 + \boldsymbol{z}_{1,ij}\boldsymbol{\Gamma}_1\boldsymbol{b}_i + \varepsilon_{ij}$ for $i = 1, ..., 1702$, and $j = 1, ..., n_i$. I let $\boldsymbol{x}_{1,ij} = (1, DBP_{ij}, SBP_{ij}, BMI_i,\ PDC_i, Gender_i,\ DBP_{ij} \times Gender_i, SBP_{ij} \times Gender_i, BMI_i \times Gender_i, PDC_i \times Gender_i,\ NumComorbid_i, NumMed_i,\ IndexAge_i)$ be the design matrix of the fixed effects

and $\boldsymbol{z}_{1,ij} = (1, DBP_{ij}, SBP_{ij}, BMI_i, PDC_i)$ be the design matrix of the random effects. I assume that $\boldsymbol{b}_i$ follows $N(0, \boldsymbol{I}_5)$ and I let $\varepsilon_{ij} \sim i.i.d.N(0, \sigma^2)$ be the measurement error.

For mortality, I assume that the survival time t_i follows a Weibull distribution. I use a proportional hazard model $h(t_i) = h_0(t_i)\exp(\boldsymbol{x}_{2,i}\boldsymbol{\beta}_2 + \boldsymbol{z}_{2,i}\boldsymbol{\Gamma}_2\boldsymbol{b}_i)$, with baseline hazard $h_0(t_i) = \alpha\lambda t_i^{\alpha-1}$ for $i = 1, ..., 1702$, where α is the shape parameter and λ is the scale parameter. I let $\boldsymbol{x}_{2,i} = (1, DBP_{i1}, SBP_{i1}, BMI_i, PDC_i, Gender_i, DBP_{i1} \times Gender_i, SBP_{i1} \times Gender_i, BMI_i \times Gender_i, PDC_i \times Gender_i, NumComorbid_i, NumMed_i, IndexAge_i)$ be the design matrix for the fixed effects and $\boldsymbol{z}_{2,i} = (1, DBP_{i1}, SBP_{i1}, BMI_i, PDC_i)$ be the design matrix for the random effects. Given the random effect $\boldsymbol{b}_i$, I assume that $\log(BNP)_{ij}$, $\log(BNP)_{ij'}$ and t_i are conditionally independent.

Data analytical results are presented in Table 2.11. For longitudinally measured BNP, our procedure selects DBP, BMI, NumComorbid, and IndexAge as non-zero fixed effects; SBP and PDC as non-zero random effects. For the survival outcome, NumMed is selected as the non-zero fixed effect; PDC as non-zero random effect. The residual plots Figure 2.1 show no violation of basic model assumptions for the two outcomes. The selected model has a smaller BIC value than the full model and a reduced model including all fixed effects and random intercept.

The effects of the selected variables on the outcomes are in expected directions. In the longitudinal model, DBP is positively associated with BNP ($\beta = 0.0145$) (greater diastolic dysfunction is associated with increased BNP level). BMI exhibits a significant negative association with BNP. For each unit of increase in BMI, log-BNP level decreases by 0.0299 ($\beta = -0.0299$). This result is not surprising as patients at advanced stage of CHF (indicated by greater BNP values) tend to have deteriorated health and much reduced body weight. Interestingly, blood pressure is not found to be associated with the survival outcome, which is influenced more strongly by the number of medications. Patients taking more medica-

Table 2.11: Results for the heart failure patient data analysis.

	Longitudinal component		Survival component	
	Fixed Effect[a]	Variance Component[b]	Fixed Effect[a]	Variance Component[b]
Intercept	5.0042±0.2321	2.6735	-	0.9657
DBP	0.0145±0.0016	0	0	0
SBP	0	0.0133	0	0
BMI	-0.0299±0.0030	0	0	0
PDC	0	2.8857	0	1.7911
Gender	0	-	0	-
DBP × Gender	0	-	0	-
SBP × Gender	0	-	0	-
BMI × Gender	0	-	0	-
PDC × Gender	0	-	0	-
Num. of comorbidities	0.1197±0.0196	-	0	-
Num. of drugs	0	-	-0.1163±0.0121	-
Index Age	-0.0033±0.0022	-	0.0044±0.0024	-

[a] Estimate of $\boldsymbol{\beta}_1 \pm SE$ and $\boldsymbol{\beta}_2 \pm SE$.
[b] Estimate of diag($\sqrt{\boldsymbol{D}_1}$) and diag($\sqrt{\boldsymbol{D}_2}$).

tions have reduced mortality risk ($\beta = -0.1163$). Patients who are older at CHF diagnosis tend to have significantly increased mortality risk ($\beta = 0.0044$). PDC, our primary variable of interest, has non-zero random effects in both longitudinal (SD=2.8857) and survival (SD=1.7911) components, which implies that medication adherence is the underlying latent process influencing both the BNP level and patient survival, and further suggests that the effects of medication adherence on the outcomes may vary across subjects. The shared random intercepts in the longitudinal component (SD=2.6735), and in the survival component (SD=0.9657) are also non-zero, which implies a strong within-patient correlation between the two outcomes as well.

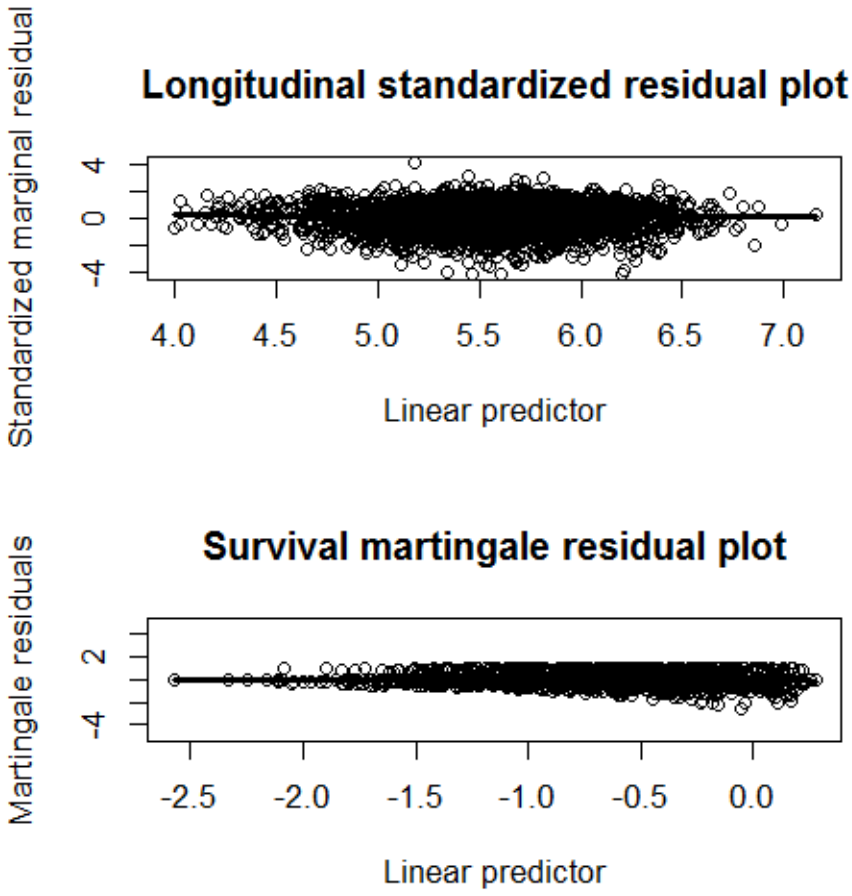

Figure 2.1: Residual plots for data application diagnostics. The circles are the standardized residuals. The black lines are the LOESS estimates.

2.5 Discussion

Despite the increasing popularity of joint models in practical data analysis, few variable selection tools are available for identifying appropriate models. In this paper, I propose a method that simultaneously selects random and fixed effects in a joint model setting. For random effect selection, I apply a Cholesky parametrization to the covariance matrix of random effects and use a group penalty, as previous studies have done (Bondell et al., 2010; Ibrahim et al., 2011). This parametrization has made the mixed-effects selection easily adaptable in the complicated joint model settings. Our simulation study shows that the proposed method could correctly identify important fixed and random effects simultaneously, even in the presence of a high proportion of censoring and a large number of random effects. The two-stage model fitting process has helped to control the estimation biases caused by the inclusion of penalty.

A major challenge of using penalized likelihood for variable selection is the computational complexity. The observed likelihood or the E-step in the EM algorithm involves analytically intractable integration. The MCMC method for integral approximation is computationally intensive. Laplace approximation could be a useful alternative, as it has been shown to offer improved computation efficiency at the expense of extra estimation bias (Ye et al., 2008a). The Gaussian quadrature method used in the current study exhibits excellent stability (At the threshold of 10^{-7}, our simulation shows a 100% convergence rate in Scenarios 2-6, and 92% convergence rate in Scenario 1; Generally, the simulation converges within 60 iterations). As I have demonstrated in the simulation, the proposed method can easily handle up to eight random effects. A possible alternative of Gaussian quadrature is the pseudo-adaptive Gauss-Hermite quadrature rule, which has been shown to be faster in computation with comparable accuracy in the joint model setting (Rizopoulos, 2012). In practice, considering the fact that most biomedical applications use random effects to

accommodate structured data dependency, thus will have a relatively small numbers of random effects, I contend that the proposed method is likely adequate for most applications. Additionally, as I have demonstrated through simulation, the number of quadrature points has limited impact on the accuracy of model selection. As a result, for complicated models one could use a smaller number of quadrature points to enhance computational efficiency in the first stage, and then increase the number of quadrature points in the second stage to achieve desired estimation accuracy. Comparing our simulation results with the reported performance in linear mixed-effects models (Bondell et al., 2010), generalized linear mixed models (Ibrahim et al., 2011), and survival models (Zhang and Lu, 2007), I note that our method has achieved comparable selection and estimation accuracy.

In summary, I show that penalized likelihood method can be used for variable selection in joint model settings. The procedure can be modified for the simultaneous mixed-effects selection in other bi-component models. Our research has demonstrated, through a real data example, that the proposed method provides a useful tool for practical data analysis. The method is easy to implement and it is efficient in computation.

Chapter 3

Structural Discovery

3.1 Introduction

Nonparametric additive models proposed by Friedman and Stuetzle (1981) and Hastie and Tibshirani (1990) can be extended to multivariate settings. These models are generally more flexible in the accommodation of potentially nonlinear effects as compared with linear models, and require fewer model assumptions, thus reducing the risk of model misspecification. Additionally, nonparametric additive models are free of the "curse of dimensionality" caused by the use of multivariate smoother. This gain in flexibility comes at the price of increased model complexity and reduced parameter interpretability. In practice, when a large number of independent variables are considered, naively assuming nonparametric effects for all independent variables in an additive structure could greatly increase the burden of parameter estimation and reduce the model efficiency. An ideal model should represent a sensible compromise between the efficiency of a linear model and the flexibility of a nonparametric model. In other words, if the true effect of an independent variable is linear, one would like to model it as such; if the true effect of an independent variable is nonlinear, one would like to model it nonparametrically; if an independent variable has no effect, one would like to exclude it from the model.

A semiparametric model combines the strength of both the linear models and nonparametric additive models. There is sizable literature on semiparametric additive models. Among the most significant developments, Martinussen and Scheike (1999) studied semiparametric additive regression models for longitudinal data. In a survival analysis setting, Huang et al. (1999) presented a semiparametric additive Cox model with polynomial splines.

More recently, Yu et al. (2012) developed a semiparametric frailty model for clustered failure time data through smoothing splines estimated by using a penalized partial likelihood method. A recent extension of the semiparametric additive model is in a joint model setting. Joint modeling of longitudinal and survival outcomes through shared latent processes provides a useful way to alleviate biases while ensuring valid inference concerning the correlation structure between the two outcomes (Faucett and Thomas, 1996). Wulfsohn and Tsiatis (1997) proposed a general framework in which the survival component was depicted by a proportional hazard model, and the longitudinal component was represented by a linear mixed-effects model. The semiparametric additive model extends the applicability of joint models. Ye et al. (2008b) incorporated smoothing spline component into joint models. Zhangsheng and Liu (2011) used a penalized spline method for joint models for recurrent and terminal events.

So far, most of the published papers have focused on the estimation of semiparametric additive models. Little has been discussed about the selection of the functional forms of the independent variables. In practice, it is of essential importance to correctly specify the functional form of an independent variable, thus serving as a safeguard for constructing a valid semiparametric model. Such a procedure is often referred to as "structural discovery" (SD). Unfortunately, the existing literature on this topic is rather limited. A naive approach to determine the functional form is to simply specify a nonparametric function for each independent effect, and then determine the functional form of each term by visualizing the independent effects. This method may be useful in some circumstances, but it lacks theoretical justifications. Another commonly used method is to test the linearity of each independent variable effect. But in complex model settings, such as in joint models, it is usually difficult to derive proper testing statistics. When there is a large number of independent variables, the power of the tests could be low and type I error rate may be

inflated due to multiple testing. More recently, a data-driven structural discovery method has been proposed by Zhang et al. (2011) in a partially linear model setting. Exhaustive literature search reveals that no similar work has been done for joint models. This may be due to the complexity of the joint model structure. The need to simultaneously perform structural discovery in both longitudinal and survival components also presents a daunting challenge as the longitudinal and survival outcomes are modeled separately through different structures.

The purpose of the current chapter is to fill in this methodology gap for nonparametric additive model in a joint model setting by using a penalized likelihood method based model selection approaches. Specifically, I propose to start from the semiparametric joint model with unspecified functional forms depicted by cubic B-splines. Then by following Wand and Ormerod (2008)'s approach, I decompose the cubic B-splines into linear and nonlinear elements. I then use variable selection methods to select the linear and nonlinear elements. This decomposition represents the model in a mixed-effect model format, which serves as a bridge to connect the structural discovery and mixed-effect selection, where the selection of linear elements mimics the fixed-effect selection, and the selection of nonlinear elements mimics the random-effect selection. Mixed-effect selection through data-driven methods has received increasing attentions in recent years. Bondell et al. (2010) studied mixed-effect selection in linear mixed-effects model. Ibrahim et al. (2011) studied mixed-effect selection in generalized linear mixed models. Most recently, He et al. (2014) studied mixed-effect selection in a joint model setting. I propose to use penalized likelihood to mimic the mixed-effect selection for structural discovery in the semiparametric additive joint models, in which a penalized likelihood will be optimized through an EM algorithm.

3.2 Method

3.2.1 Model Formulation

To perform structural discovery in a semiparametric additive joint model, I introduce a semiparametric linear mixed-effects model for the longitudinal component and a semiparametric frailty model for the survival component, in which the functional forms of continuous variables with unknown effects are modeled by cubic B-splines.

Suppose in a longitudinal study, I have a survival outcome (t_i, δ_i), and repeated measurements of a continuous outcome $\boldsymbol{y}_i$ for $i = 1, \cdots, n$ subjects. Here t_i is the observed event time subjected to right censoring, and δ_i is a failure indicator with $\delta_i = 1$ indicating the occurrence of an event of interest, and $\delta_i = 0$ indicating censoring, whereas $\boldsymbol{y}_i = \{y_{i1} \dots y_{in_i}\}$ is an $n_i \times 1$ vector of the n_i repeated measurements. For the longitudinal component, let $\boldsymbol{X_{1i}} \in \mathbb{R}^{n_i \times p}$ be the covariate matrix for unknown effects with jth row vector $\boldsymbol{x_{1ij}} \in \mathbb{R}_1^p$ and $\boldsymbol{W_{1i}} \in \mathbb{R}^{n_i \times d}$ be the covariate matrix for effects already identified as linear with jth row vector $\boldsymbol{w_{1ij}} \in \mathbb{R}_1^d$. For the survival component, let $\boldsymbol{x_{2i}} \in \mathbb{R}_1^p$ and $\boldsymbol{w_{2i}} \in \mathbb{R}_1^d$ be the covariate vectors for unknown effects and known linear effects, respectively. I denote $\boldsymbol{Z_{1i}} \in \mathbb{R}^{n_i \times q}$ and $\boldsymbol{z_{2i}} \in \mathbb{R}_1^q$ as the random effect covariate matrix and vector in the two model components. Without loss of generality, I let the longitudinal and survival components have the same set of covariates and random effects here. This could be easily generalized to the situation with different sets of covariates and random effects in each component. Combining these observations I write $\boldsymbol{O_i} = (\boldsymbol{y}_i, \boldsymbol{X}_{1i}, \boldsymbol{W_{1i}}, \boldsymbol{Z}_{1i}, t_i, \delta_i, \boldsymbol{x}_{2i}, \boldsymbol{w_{2i}}, \boldsymbol{z}_{2i})$. I assume that the observations $\boldsymbol{O_i}$ are independent across subjects.

Then, for the longitudinal component, I assume a semiparametric linear mixed-effects model:

$$y_{ij} = \beta_0 + \sum_{h=1}^{p} f_h(x_{1ij,h}) + \boldsymbol{w_{1ij}^T \gamma_1} + \boldsymbol{z_{1ij}^T b_i} + \varepsilon_{ij}, \tag{3.1}$$

and for the survival component, I assume a semiparametric frailty model:

$$h(t_i) = h_0(t_i) \exp\{\sum_{h=1}^{p} g_h(x_{2i,h}) + \boldsymbol{w_{2i}^T \gamma_2} + \boldsymbol{z_{2i}^T b_i}\}, \tag{3.2}$$

where $x_{1ij,h}$ and $x_{2i,h}$ are the hth elements in the independent variable vectors $\boldsymbol{x_{1ij}}$ and $\boldsymbol{x_{2i}}$; $f_h(\cdot)$ and $g_h(\cdot)$ are the corresponding unknown cubic B-spline functions; $\boldsymbol{\gamma_1}$ and $\boldsymbol{\gamma_2}$ are regression coefficients for linear effects; $\boldsymbol{b_i} = (b_1, \cdots, b_q) \in \mathbb{R}_1^q$ is a random effect vector following a multivariate normal distribution $MVN(\boldsymbol{0, D(\phi)})$; $\varepsilon_{ij} \sim N(0, \sigma^2)$ is the measurement error, which is assumed to be independently and identically distributed; β_0 is the intercept of regression function and $h_0(t_i)$ is the baseline hazard function. The baseline hazard function $h_0(t_i)$ could be modeled nonparametrically through B-spline. One could also parameterize $h_0(t_i)$ with a parametric distribution. For example, a natural option is to use a Weibull distribution with a baseline hazard $h_0(t_i) = \alpha\lambda t_i^{\alpha-1}$, where α is the shape parameter and λ is the scale parameter. Alternatively, one could use a piece-wise constant baseline hazard by dividing the study period into m intervals and assuming $h_0(t)$ to be a constant on each interval as

$$h_0(t) = h_k, t_{k-1} < t \leq t_k, k = 1 \ldots m, \tag{3.3}$$

where t_ks are knots that define the intervals. This piece-wise constant baseline hazard had been shown to perform well by Feng et al. (2005) and Ding and Wang (2008).

Let $\boldsymbol{\eta} = \{f_h(\cdot), g_h(\cdot), \boldsymbol{\gamma_1}, \boldsymbol{\gamma_2}, \boldsymbol{\phi}, \sigma, h_0(\cdot)\}$ be the vector of all unknown parameters. The marginal likelihood of $\boldsymbol{\eta}$ could be written as:

$$\begin{aligned} L_o(\boldsymbol{\eta}) &= \prod_{i=1}^{n} \int f(\boldsymbol{y_i}, t_i, \delta_i | \boldsymbol{b_i}, \boldsymbol{X}_{1i}, \boldsymbol{W}_{1i}, \boldsymbol{Z}_{1i}, \boldsymbol{x}_{2i}, \boldsymbol{w}_{2i}, \boldsymbol{z}_{2i}, \boldsymbol{\eta}) f_b(\boldsymbol{b}_i) d\boldsymbol{b}_i \\ &= \prod_{i=1}^{n} \int \prod_{j=1}^{n_i} f_y(\boldsymbol{y}_{ij} | \boldsymbol{x}_{1ij}, \boldsymbol{w}_{1ij}, \boldsymbol{z}_{1ij}, \boldsymbol{b}_i, \boldsymbol{\eta}) f_s(t_i, \delta_i | \boldsymbol{x}_{2i}, \boldsymbol{w}_{2i}, \boldsymbol{z}_{2i}, \boldsymbol{b}_i, \boldsymbol{\eta}) f_b(\boldsymbol{b}_i) d\boldsymbol{b}_i, \end{aligned} \tag{3.4}$$

where $f_b(\boldsymbol{\cdot})$ is a $q-$variate normal density function for $\boldsymbol{b_i}$, $f_s(\cdot)$ is the likelihood of the survival component parameters conditional on $\boldsymbol{b_i}$, and $f_y(\cdot)$ is the density function of repeated measurements conditional on $\boldsymbol{b_i}$.

3.2.2 Penalized Smoothing Splines

Penalized cubic B-spline is used to estimate $f_h(\cdot)$ and $g_h(\cdot)$. To illustrate penalized cubic B-spline, I use $f(\cdot)$ as an example. For simplicity, I omit the subscription. Let $B_1, \ldots, B_{K+4}$ be the cubic B-spline basis functions with K inner knots for $f(x)$, and suppose we have n observations $(x_1, \ldots, x_n)$, thus we could set up the $n \times (K+4)$ design matrix $\boldsymbol{B}$, with row vector as $(B_1(x_i), \ldots, B_{K+4}(x_i))$. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{K+4})^T$ be the corresponding B-spline regression coefficients. Then $\boldsymbol{B}\widehat{\boldsymbol{\theta}}$ can be used to estimate $f(x)$. The penalty term could be written as: $\boldsymbol{\theta^T \Omega \theta}$, where $\boldsymbol{\Omega}_{(K+4)\times(K+4)}$ is the penalty matrix with $\Omega_{kk'} = \int B_k^{(2)}(s) B_{k'}^{(2)}(s) ds$. Then the penalized likelihood for estimating $f_h(\cdot)$ and $g_h(\cdot)$ can be written as

$$pl_{original} = l_o(\boldsymbol{\eta}) - \lambda_1 \sum_{h=1}^{p} \boldsymbol{\theta}_{\boldsymbol{f,h}}^{\boldsymbol{T}} \boldsymbol{\Omega}_{\boldsymbol{f,h}} \boldsymbol{\theta}_{\boldsymbol{f,h}} - \lambda_2 \sum_{h=1}^{p} \boldsymbol{\theta}_{\boldsymbol{g,h}}^{\boldsymbol{T}} \boldsymbol{\Omega}_{\boldsymbol{g,h}} \boldsymbol{\theta}_{\boldsymbol{g,h}} \tag{3.5}$$

where $l_o(\cdot) = \log L_o(\cdot)$ and $L_o(\cdot)$ is defined in (3.4), and λ_1 and λ_2 are the smoothing parameters controlling the goodness of fit and the smoothness of the curves. Derivation for each of $f_h(\cdot)$s and $g_h(\cdot)$s can be done in the same way.

3.2.3 Structural Discovery Using Reparametrized Penalized Smoothing Splines

The penalized likelihood in (3.5) only estimates the cubic B-spline functions $f_h(\cdot)$s and $g_h(\cdot)$s. It does not by itself perform the structural discovery for identifying the true effects as depicted by $f_h(\cdot)$s and $g_h(\cdot)$s. To discover the true effects, I decompose $f_h(\cdot)$s and $g_h(\cdot)$s into linear and nonlinear elements, and then use a variable selection method to select these elements.

Following Wand and Ormerod (2008), by spectral decomposition, I decompose the penalty matrix $\boldsymbol{\Omega}$ as $\boldsymbol{\Omega} = \boldsymbol{U}\text{diag}(\boldsymbol{d})\boldsymbol{U}^{\boldsymbol{T}}$. Matrix $\boldsymbol{U}$ consists of column eigenvectors and vector $\boldsymbol{d}$ consists of eigenvalues arranged in descending order. Let $\boldsymbol{d} = (\boldsymbol{d}_+^{\boldsymbol{T}}, \boldsymbol{d}_{\boldsymbol{0}}^{\boldsymbol{T}})^{\boldsymbol{T}}$, where $\boldsymbol{d}_+^{\boldsymbol{T}}$ is the vector of $K+2$ descending positive eigenvalues, and $\boldsymbol{d}_{\boldsymbol{0}}^{\boldsymbol{T}}$ is the vector of two zero eigenvalues. Let $\boldsymbol{U} = [\boldsymbol{U}_+, \boldsymbol{U}_{\boldsymbol{0}}]$, where $\boldsymbol{U}_+$ is a matrix of dimension $(K+4) \times (K+2)$, corresponding to $\boldsymbol{d}_+$, and $\boldsymbol{U}_{\boldsymbol{0}}$ is a matrix of dimension $(K+4) \times 2$, corresponding to $\boldsymbol{d}_{\boldsymbol{0}}$,

Under this decomposition, cubic B-spline could be written as:

$$\begin{aligned}
\boldsymbol{B\theta} &= \boldsymbol{BUU}^{\boldsymbol{T}}\boldsymbol{\theta} \\
&= \boldsymbol{B}[\boldsymbol{U}_{\boldsymbol{0}}\boldsymbol{U}_{\boldsymbol{0}}^{\boldsymbol{T}}\boldsymbol{\theta} + \boldsymbol{U}_+\text{diag}(\boldsymbol{d}_+^{-\boldsymbol{1/2}})\text{diag}(\boldsymbol{d}_+^{\boldsymbol{1/2}})\boldsymbol{U}_+^{\boldsymbol{T}}\boldsymbol{\theta}] \\
&= \boldsymbol{B}[\boldsymbol{U}_{\boldsymbol{0}}\boldsymbol{\beta} + \boldsymbol{U}_+\text{diag}(\boldsymbol{d}_+^{-\boldsymbol{1/2}})\boldsymbol{u}] \\
&= \boldsymbol{C\beta} + \boldsymbol{Au},
\end{aligned} \tag{3.6}$$

where $\boldsymbol{C} = \boldsymbol{BU}_{\boldsymbol{0}}, \boldsymbol{\beta} = \boldsymbol{U}_{\boldsymbol{0}}^{\boldsymbol{T}}\boldsymbol{\theta}$, $\boldsymbol{A} = \boldsymbol{BU}_+\text{diag}(\boldsymbol{d}_+^{-\boldsymbol{1/2}})$, and $\boldsymbol{u} = \text{diag}(\boldsymbol{d}_+^{\boldsymbol{1/2}})\boldsymbol{U}_+^{\boldsymbol{T}}\boldsymbol{\theta}$. The penalty term can be written as:

$$\begin{aligned}
\boldsymbol{\theta}^{\boldsymbol{T}}\boldsymbol{\Omega\theta} &= \boldsymbol{\theta}^{\boldsymbol{T}}\boldsymbol{U}\text{diag}(\boldsymbol{d})\boldsymbol{U}^{\boldsymbol{T}}\boldsymbol{\theta} \\
&= \boldsymbol{\theta}^{\boldsymbol{T}}\boldsymbol{U}_{\boldsymbol{0}}\text{diag}(\boldsymbol{d}_{\boldsymbol{0}})\boldsymbol{U}_{\boldsymbol{0}}^{\boldsymbol{T}}\boldsymbol{\theta} + \boldsymbol{\theta}^{\boldsymbol{T}}\boldsymbol{U}_+\text{diag}(\boldsymbol{d}_+)\boldsymbol{U}_+^{\boldsymbol{T}}\boldsymbol{\theta} \\
&= \boldsymbol{u}^{\boldsymbol{T}}\boldsymbol{u}.
\end{aligned} \tag{3.7}$$

Thus the decomposition represents the model in a mixed-effect model format, in which $\boldsymbol{C\beta}$ mimics the fixed effect and $\boldsymbol{Au}$ mimics the random effect. As in Wand and Ormerod (2008), $\boldsymbol{C}$ is the basis for the linear space and $\boldsymbol{A}$ is the basis for nonlinear effect, and $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are the corresponding regression coefficients. Thus selection of linear and nonlinear effects mimics the seletion of mixed effects. A simple specification of the linear basis is $\boldsymbol{C} = [1, x_i]$ (Speed, 1991; Wand and Ormerod, 2008). To ensure identifiability, I omit the intercept in $\boldsymbol{C} = [1, x_i]$ and centralize x_i by following Yu et al.,(2012).

With this reparameterization, I re-write the nonlinear functions in Equation (3.1) as $f_h(x_{1ij,h}) = C_{1ij,h}\beta_{1h} + \boldsymbol{A_{1ij,h}u_{1h}}$ and the nonlinear functions in Equation (3.2) as $g_h(x_{2i,h}) = C_{2i,h}\beta_{2h} + \boldsymbol{A_{2i,h}u_{2h}}$. The reparameterized semiparametric linear mixed-effects model therefore takes the following form:

$$y_{ij} = \beta_0 + \sum_{h=1}^{p}(C_{1ij,h}\beta_{1h} + \boldsymbol{A_{1ij,h}u_{1h}}) + \boldsymbol{w_{1ij}^T\gamma_1} + \boldsymbol{z_{1ij}^T b_i} + \varepsilon_{ij}, \tag{3.8}$$

and the reparameterized semiparametric frailty model can be written as:

$$h(t_i) = h_0(t_i)exp\{\sum_{h=1}^{p}(C_{2i,h}\beta_{2h} + \boldsymbol{A_{2i,h}u_{2h}}) + \boldsymbol{w_{2i}^T\gamma_2} + \boldsymbol{z_{2i}^T b_i}\}, \tag{3.9}$$

where $C_{2i,h} = (x_{2i,h} - \bar{x}^0_{2h})$ and $\bar{x}^0_{2h} = \sum_{k=1}^{r_h} x^0_{2hk}/r_h$, where x^0_{2hk} $(k = 1, \ldots, r_h)$ are the ordered r_h distinct independent variable values for the hth independent variable x_{2h}. Similarly, $C_{1ij,h} = (x_{1ij,h} - \bar{x}^0_{1h})$.

Let $\boldsymbol{\zeta} = (\beta_0, \beta_{1h}, \beta_{2h}, \boldsymbol{u_{1h}}, \boldsymbol{u_{2h}}, \boldsymbol{\gamma_1}, \boldsymbol{\gamma_2}, \boldsymbol{\phi}, \sigma, h_0(t_i))$, then $pl_{original}$ in (3.5) could be written as the smoothing spline analysis of variance (SSANOVA) type penalized likelihood

$$pl_{SSANOVA} = l_o(\boldsymbol{\zeta}) - \lambda_1 \sum_{h=1}^{p} \boldsymbol{u_{1h}^T u_{1h}} - \lambda_2 \sum_{h=1}^{p} \boldsymbol{u_{2h}^T u_{2h}}. \tag{3.10}$$

However, the penalty terms of $pl_{SSANOVA}$ in Equation (3.10) do not possess the sparsity property for variable selection of regression coefficients $\boldsymbol{u_{1h}}$ and $\boldsymbol{u_{2h}}$ for the nonlinear effects. In addition, the penalized likelihood $pl_{SSANOVA}$ does not allow the selection of regression coefficients β_{1h} and β_{2h} for linear effects. To perform structural discovery, I propose the following penalized likelihood pl_{SD} with penalty terms having sparsity on estimation of β_{1h}, β_{2h}, $\boldsymbol{u_{1h}}$ and $\boldsymbol{u_{2h}}$:

$$\begin{aligned} pl_{SD}(\boldsymbol{\zeta}) = \frac{1}{n} l_o(\boldsymbol{\zeta}) - \lambda_{1,\beta} \sum_{h=1}^{p} \kappa_{\lambda_{1,\beta}}(\beta_{1h}) - \lambda_{2,\beta} \sum_{h=1}^{p} \kappa_{\lambda_{2,\beta}}(\beta_{2h}) \\ - \lambda_{1,\boldsymbol{u}} \sum_{h=1}^{p} \kappa_{\lambda_{1,\boldsymbol{u}}}(\boldsymbol{u_{1h}}) - \lambda_{2,\boldsymbol{u}} \sum_{h=1}^{p} \kappa_{\lambda_{2,\boldsymbol{u}}}(\boldsymbol{u_{2h}}). \end{aligned} \tag{3.11}$$

The penalty functions $\kappa_{\lambda_{1,\beta}}(\beta_{1h})$ and $\kappa_{\lambda_{2,\beta}}(\beta_{2h})$ control the sparsity of estimates of β_{1h} and β_{2h} so that the linear effects are selected, where $\lambda_{1,\beta}$ and $\lambda_{2,\beta}$ are the associated positive tuning parameters. The penalty terms $\kappa_{\lambda_{1,\boldsymbol{u}}}(\boldsymbol{u_{1h}})$ and $\kappa_{\lambda_{2,\boldsymbol{u}}}(\boldsymbol{u_{2h}})$ control the sparsity of estimates of $\boldsymbol{u_{1h}}$ and $\boldsymbol{u_{2h}}$ for selection of the nonlinear effects, where $\lambda_{1,\boldsymbol{u}}$ and $\lambda_{2,\boldsymbol{u}}$ are the associated positive tuning parameters; $\kappa_{\lambda_{1,\beta}}(\beta_{1h})$ and $\kappa_{\lambda_{1,\boldsymbol{u}}}(\boldsymbol{u_{1h}})$ jointly determine the effect of hth independent variable x_{1h} in the longitudinal component; $\kappa_{\lambda_{2,\beta}}(\beta_{2h})$ and $\kappa_{\lambda_{2,\boldsymbol{u}}}(\boldsymbol{u_{2h}})$ jointly determine the effect of hth independent variable x_{2h} in the survival component. Here, I define the partially linear effect as β and $\boldsymbol{u}$ are both nonzero. The functions $\kappa_{\lambda_{1,\beta}}(\cdot)$, $\kappa_{\lambda_{2,\beta}}(\cdot)$, $\kappa_{\lambda_{1,\boldsymbol{u}}}(\cdot)$, $\kappa_{\lambda_{2,\boldsymbol{u}}}(\cdot)$ could be the adaptive LASSO or the smoothly clipped absolute deviation (SCAD) penalties.

For linear-effect selection, I define the adaptive LASSO penalty as $\kappa_{\lambda_{1,\beta}}(\beta_{1h}) = \omega_{\beta_{1h}}|\beta_{1h}|$ and $\lambda_{2,\beta_{2h}} = \omega_{\beta_{2h}}|\beta_{2h}|$, where $\omega_{\beta_{1h}}, \omega_{\beta_{2h}}$ are the corresponding positive weights for penalty $|\beta_{1h}|$ and $|\beta_{2h}|$. I choose the weights as $\omega_{\beta_{1h}} = |\tilde{\beta}_{1h}|^{-1}, \omega_{\beta_{2h}} = |\tilde{\beta}_{2h}|^{-1}$, where $\tilde{\beta}_{1h}$ and $\tilde{\beta}_{2h}$ are the optimizers of the SSANOVA type penalized likelihood defined in Equation (3.10).

Some of the estimates of $\hat{\beta}_{1h}$ and $\hat{\beta}_{2h}$ for penalized likelihood (3.11) will be zero since $|\beta_{1h}|$ and $|\beta_{2h}|$ are singular when $|\beta_{1h}| = 0$ and $|\beta_{2h}| = 0$.

For nonlinear effect selection, I note that a nonlinear effect could be excluded if and only if $\boldsymbol{u_{1h}} = 0$ and $\boldsymbol{u_{2h}} = 0$. As $\boldsymbol{u_{1h}}$ and $\boldsymbol{u_{2h}}$ are vectors, I propose to perform the group penalty on $\boldsymbol{u_{1h}}$ and $\boldsymbol{u_{2h}}$. I first summarize it using an L_2-norm: $||\boldsymbol{u_{1h}}|| = (\boldsymbol{u_{1h}^T u_{1h}})^{1/2}$ and $||\boldsymbol{u_{2h}}|| = (\boldsymbol{u_{2h}^T u_{2h}})^{1/2}$. Following Yuan and Lin (2006), I define the adaptive LASSO penalties for nonlinear effect as: $\kappa_{\lambda_{1,\boldsymbol{u}}}(\boldsymbol{u_{1h}}) = \omega_{\boldsymbol{u_{1h}}}||\boldsymbol{u_{1h}}||$ and $\kappa_{\lambda_{2,\boldsymbol{u}}}(\boldsymbol{u_{2h}}) = \omega_{\boldsymbol{u_{2h}}}||\boldsymbol{u_{2h}}||$. The weights are chosen as $\omega_{\boldsymbol{u_{1h}}} = ||\tilde{\boldsymbol{u}}_{1h}||^{-1}, \omega_{\boldsymbol{u_{2h}}} = ||\tilde{\boldsymbol{u}}_{2h}||^{-1}$, where $\tilde{\boldsymbol{u}}_{1h}$ and $\tilde{\boldsymbol{u}}_{2h}$ are the optimizers of the SSANOVA type penalized likelihood defined in Equation (3.10). The penalized likelihood with the SCAD penalty terms could be constructed in a similar way by substituting the penalty term in (3.11) with SCAD penalty on $(|\beta_{1h}|, |\beta_{2h}|, ||\boldsymbol{u_{1h}}||, ||\boldsymbol{u_{2h}}||)$. The estimator of $\boldsymbol{\zeta} = (\beta_0, \beta_{1h}, \beta_{2h}, \boldsymbol{u_{1h}}, \boldsymbol{u_{2h}}, \boldsymbol{\gamma_1}, \boldsymbol{\gamma_2}, \boldsymbol{\phi}, \sigma, h_0(t_i))$ can be obtained by maximizing Equation (3.11). I use adaptive LASSO penalty in the simulation study.

3.2.4 EM Algorithm for Optimization of the Penalized Likelihood

To maximize the penalty likelihood (3.11), I use an EM algorithm. Let

$$\begin{aligned} p(\boldsymbol{\zeta}) =& \lambda_{1,\beta} \sum_{h=1}^{p} \omega_{\beta_{1h}} |\beta_{1h}| + \lambda_{2,\beta} \sum_{h=1}^{p} \omega_{\beta_{2h}} |\beta_{2h}| + \\ & \lambda_{1,\boldsymbol{u}} \sum_{h=1}^{p} \omega_{\boldsymbol{u_{1h}}} ||\boldsymbol{u_{1h}}|| + \lambda_{2,\boldsymbol{u}} \sum_{h=1}^{p} \omega_{\boldsymbol{u_{2h}}} ||\boldsymbol{u_{2h}}|| \end{aligned}$$

denote the penalty terms. I start with the penalized complete likelihood for ($\boldsymbol{\zeta}$) for $i = 1, \cdots, n$, which is

$$\begin{aligned} pl_c(\boldsymbol{\zeta}) =& \frac{1}{n}\sum_{i=1}^{n} \log f(\boldsymbol{y}_i, t_i, \delta_i, \boldsymbol{b}_i|\boldsymbol{\zeta}) \\ & - \lambda_{1,\beta}\sum_{h=1}^{p}\kappa_{\lambda_{1,\beta}}(\beta_{1h}) - \lambda_{2,\beta}\sum_{h=1}^{p}\kappa_{\lambda_{2,\beta}}(\beta_{2h}) \\ & - \lambda_{1,\boldsymbol{u}}\sum_{h=1}^{p}\kappa_{\lambda_{1,\boldsymbol{u}}}(\boldsymbol{u_{1h}}) - \lambda_{2,\boldsymbol{u}}\sum_{h=1}^{p}\kappa_{\lambda_{2,\boldsymbol{u}}}(\boldsymbol{u_{2h}}) \\ =& \frac{1}{n}\sum_{i=1}^{n}\{\log[f_y(\boldsymbol{y_i}|\boldsymbol{b}_i, \boldsymbol{\zeta})] + \delta_i \log[h(t_i|\boldsymbol{b}_i, \boldsymbol{\zeta})] + \log[S(t_i|\boldsymbol{b}_i, \boldsymbol{\zeta})] + \log[f_b(\boldsymbol{b}_i|\boldsymbol{\zeta})]\} - p(\boldsymbol{\zeta}). \end{aligned} \tag{3.12}$$

In Equation (3.12), $S(\cdot)$ is the survival function of t_i conditional on $\boldsymbol{b_i}$. Let $\boldsymbol{\lambda} = (\lambda_{1,\beta}, \lambda_{2,\beta}, \lambda_{1,\boldsymbol{u}}, \lambda_{2,\boldsymbol{u}})$ and $\boldsymbol{\omega} = (\omega_{\beta_{1h}}, \omega_{\beta_{2h}}, \omega_{\boldsymbol{u_{1h}}}, \omega_{\boldsymbol{u_{2h}}})$. Here I denote the complete data as $\boldsymbol{g}_{c,i} = (\boldsymbol{y}_i, \boldsymbol{X}_{1i}, \boldsymbol{x}_{2i}, t_i, \delta_i, \boldsymbol{W}_{1i}, \boldsymbol{w}_{2i}, \boldsymbol{Z}_{1i}, \boldsymbol{z}_{2i}, \boldsymbol{b}_i)$, the complete longitudinal data as $\boldsymbol{g}_{c,i1} = (\boldsymbol{y}_i, \boldsymbol{X}_{1i}, \boldsymbol{Z}_{1i}, \boldsymbol{W}_{1i}, \boldsymbol{b}_i)$, the complete survival data as $\boldsymbol{g}_{c,i2} = (t_i, \delta_i, \boldsymbol{x}_{2i}, \boldsymbol{z}_{2i}, \boldsymbol{w}_{2i}, \boldsymbol{b}_i)$, the observed data as $\boldsymbol{g}_{o,i} = (\boldsymbol{y}_i, \boldsymbol{X}_{1i}, \boldsymbol{x}_{2i}, t_i, \delta_i, \boldsymbol{W}_{1i}, \boldsymbol{w}_{2i}, \boldsymbol{Z}_{1i}, \boldsymbol{z}_{2i})$, the observed longitudinal data as $\boldsymbol{g}_{o,i1} = (\boldsymbol{y}_i, \boldsymbol{X}_{1i}, \boldsymbol{W}_{1i}, \boldsymbol{Z}_{1i})$, and the observed survival data as $\boldsymbol{g}_{o,i2} = (\boldsymbol{x}_{2i}, t_i, \delta_i, \boldsymbol{w}_{2i}, \boldsymbol{z}_{2i})$.

E-step

I first derive the E-step of the EM algorithm for the fixed tuning parameter and weights $\boldsymbol{\lambda}, \boldsymbol{\omega}$. Letting $\boldsymbol{\zeta}^{(s)}$ be from the sth iteration of maximization, I take the expectation of the penalized log-complete likelihood conditional on $\boldsymbol{\zeta}^{(s)}$ and $\boldsymbol{g}_{oi}$, for $i = 1, \ldots, n$, and obtain

the following penalized Q-function:

$$Q_{\lambda,\omega}(\tilde{\boldsymbol{\eta}}|\boldsymbol{\zeta}^{(s)}) = \frac{1}{n}\sum_{i=1}^{n}\{E[\log f_y(\boldsymbol{g}_{c,i1},\boldsymbol{\zeta})|(\boldsymbol{g}_{o,i},\boldsymbol{\zeta}^{(s)})] + E[\delta_i \log h(\boldsymbol{g}_{c,i2},\boldsymbol{\zeta})|(\boldsymbol{g}_{o,i},\boldsymbol{\zeta}^{(s)})] + E[\log S(\boldsymbol{g}_{c,i2},\boldsymbol{\zeta})|(\boldsymbol{g}_{o,i},\boldsymbol{\zeta}^{(s)})]\} - p(\boldsymbol{\zeta}) + \frac{1}{n}\sum_{i=1}^{n} E[\log f_b(\boldsymbol{b}_i)|(\boldsymbol{g}_{o,i},\boldsymbol{\zeta}^{(s)})], \tag{3.13}$$

where

$$E[H(\boldsymbol{b_i})|(\boldsymbol{g}_{o,i},\boldsymbol{\zeta}^{(s)})] = \int H(\boldsymbol{b_i}) f_b(\boldsymbol{b}_i|\boldsymbol{g}_{o,i},\boldsymbol{\zeta}^{(s)}) d\boldsymbol{b}_i, \tag{3.14}$$

for each of $H(\boldsymbol{b_i}) = \log f_y(\boldsymbol{g}_{c,i1},\boldsymbol{\zeta})$, $H(\boldsymbol{b_i}) = \delta_i \log h(\boldsymbol{g}_{c,i2},\boldsymbol{\zeta})$, $H(\boldsymbol{b_i}) = \log S(\boldsymbol{g}_{c,i2},\boldsymbol{\zeta})$ and $H(\boldsymbol{b_i}) = \log f_b(\boldsymbol{b}_i)$. Because the integral in Equation (3.14) is intractable, I approximate it by using a multivariate Gaussian quadrature (Pinheiro and Bates, 1995). Since $\boldsymbol{b_i} \in \mathbb{R}_1^q$, if I choose k quadrature points, there will be a total of k^q vector nodes of $q \times 1$ dimension. Let $\boldsymbol{b}_l^{'} = (b_{l,1}^{'}, b_{l,2}^{'}, \cdots, b_{l,q}^{'})$ denote the lth node, and w_l denote the corresponding quadrature weight, for $l = 1, \cdots, k^q$, the integral (3.14) can be approximated by:

$$\tilde{E}\{H(\boldsymbol{b_i})|(\boldsymbol{g}_{o,i},\boldsymbol{\zeta}^{(s)})\} \approx \sum_{l=1}^{k^q} w_l H(\boldsymbol{b_l'}) f_b(\boldsymbol{b_l'}|\boldsymbol{g}_{o,i},\boldsymbol{\zeta}^{(s)}) \tag{3.15}$$

Therefore, I obtain the approximated expected function to be maximized in the $(s+1)$th iteration as

$$\tilde{Q}_{\lambda,\omega}(\boldsymbol{\zeta}|\boldsymbol{\zeta}^{(s)}) = \frac{1}{n}\sum_{i=1}^{n}\{\tilde{E}[\log f_y(\boldsymbol{g}_{c,i1},\boldsymbol{\zeta})|(\boldsymbol{g}_{o,i},\boldsymbol{\zeta}^{(s)})] + \tilde{E}[\delta_i \log h(\boldsymbol{g}_{c,i2},\boldsymbol{\zeta})|(\boldsymbol{g}_{o,i},\boldsymbol{\zeta}^{(s)})] + \tilde{E}[\log S(\boldsymbol{g}_{c,i2},\boldsymbol{\zeta})|(\boldsymbol{g}_{o,i},\boldsymbol{\zeta}^{(s)})]\} - p(\boldsymbol{\zeta}) + \frac{1}{n}\sum_{i=1}^{n} \tilde{E}[\log f_b(\boldsymbol{b}_i)|(\boldsymbol{g}_{o,i},\boldsymbol{\zeta}^{(s)})] \tag{3.16}$$

M-step

I maximize (3.16) with respect to the (β_1, β_2)s and $(\boldsymbol{u_1}, \boldsymbol{u_2})$s alternatively. When $(\beta_0, \boldsymbol{u_{1h}}, \boldsymbol{u_{2h}}, \boldsymbol{\gamma_1}, \boldsymbol{\gamma_2}, \boldsymbol{\phi}, \sigma, h_0(t_i))$ are fixed, I maximize (3.16) with respect to (β_{1h}, β_{2h}), and the penalty function involving L_1 penalty terms can be solved by applying the Least Angle Regression (LARS)/LASSO algorithm (Efron et al., 2004) and the SCAD penalty could be solved according to Fan and Li (2001). When $(\beta_0, \beta_{1h}, \beta_{2h}, \boldsymbol{\gamma_1}, \boldsymbol{\gamma_2}, \boldsymbol{\phi}, \sigma, h_0(t_i))$ are fixed, I maximize (3.16) with respect to $(\boldsymbol{u_{1h}}, \boldsymbol{u_{2h}})$.

The expectation conditional maximization procedures are proposed as follows:

1. Initialize $(\beta_0^{(0)}, \beta_{1h}^{(0)}, \beta_{2h}^{(0)}, \boldsymbol{u_{1h}^{(0)}}, \boldsymbol{u_{2h}^{(0)}}, \boldsymbol{\gamma_1^{(0)}}, \boldsymbol{\gamma_2^{(0)}}, \boldsymbol{\phi^{(0)}}, \sigma^{(0)}, h_0(t_i)^{(0)})$ with some plausible values.

2. For iteration s, update β_{1h}, β_{2h} by adaptive LASSO,

$$\begin{aligned}\beta_{1h}^{(s)}, \beta_{2h}^{(s)} = \underset{\beta_{1h}, \beta_{2h}}{argmax}\, \tilde{Q}_{\lambda,\omega}(\beta_0^{(s-1)}, \beta_{1h}, \beta_{2h}, \boldsymbol{u_{1h}^{(s-1)}}, \boldsymbol{u_{2h}^{(s-1)}}, \boldsymbol{\gamma_1^{(s-1)}}, \boldsymbol{\gamma_2^{(s-1)}}, \boldsymbol{\phi^{(s-1)}}, \sigma^{(s-1)}, \\ h_0(t_i)^{(s-1)} | \beta_0^{(s-1)}, \beta_{1h}^{(s-1)}, \beta_{2h}^{(s-1)}, \boldsymbol{u_{1h}^{(s-1)}}, \boldsymbol{u_{2h}^{(s-1)}}, \boldsymbol{\gamma_1^{(s-1)}}, \boldsymbol{\gamma_2^{(s-1)}}, \\ \boldsymbol{\phi^{(s-1)}}, \sigma^{(s-1)}, h_0(t_i)^{(s-1)})\end{aligned}$$

3. Update $\boldsymbol{u_{1h}}, \boldsymbol{u_{2h}}$,

$$\begin{aligned}\boldsymbol{u_{1h}^{(s)}}, \boldsymbol{u_{2h}^{(s)}} = \underset{\boldsymbol{u_{1h}}, \boldsymbol{u_{2h}}}{argmax}\, \tilde{Q}_{\lambda,\omega}(\beta_0^{(s-1)}, \beta_{1h}^{(s)}, \beta_{2h}^{(s)}, \boldsymbol{u_{1h}}, \boldsymbol{u_{2h}}, \boldsymbol{\gamma_1^{(s-1)}}, \boldsymbol{\gamma_2^{(s-1)}}, \boldsymbol{\phi^{(s-1)}}, \sigma^{(s-1)}, \\ h_0(t_i)^{(s-1)} | \beta_0^{(s-1)}, \beta_{1h}^{(s)}, \beta_{2h}^{(s)}, \boldsymbol{u_{1h}^{(s-1)}}, \boldsymbol{u_{2h}^{(s-1)}}, \boldsymbol{\gamma_1^{(s-1)}}, \boldsymbol{\gamma_2^{(s-1)}}, \\ \boldsymbol{\phi^{(s-1)}}, \sigma^{(s-1)}, h_0(t_i)^{(s-1)})\end{aligned}$$

4. Update $\beta_0, \boldsymbol{\gamma_1}, \boldsymbol{\gamma_2}, \sigma_\varepsilon, h_0(t_i), \phi$:

$$\beta_0^{(s)}, \boldsymbol{\gamma_1^{(s)}}, \boldsymbol{\gamma_2^{(s)}}, \boldsymbol{\phi^{(s)}}, \sigma^{(s)}, h_0(t_i)^{(s)} = \underset{\beta_0, \boldsymbol{\gamma_1}, \boldsymbol{\gamma_2}, \boldsymbol{\phi}, \sigma, h_0(t_i)}{argmax} \tilde{Q}_{\lambda,\omega}(\beta_0, \beta_{1h}^{(s)}, \beta_{2h}^{(s)}, \boldsymbol{u_{1h}^{(s)}}, \boldsymbol{u_{2h}^{(s)}}, \boldsymbol{\gamma_1}, \boldsymbol{\gamma_2}, \boldsymbol{\phi}, \sigma, h_0(t_i) | \beta_0^{(s-1)}, \beta_{1h}^{(s)}, \beta_{2h}^{(s)}, \boldsymbol{u_{1h}^{(s)}}, \boldsymbol{u_{2h}^{(s)}}, \boldsymbol{\gamma_1^{(s-1)}}, \boldsymbol{\gamma_2^{(s-1)}}, \boldsymbol{\phi^{(s-1)}}, \sigma^{(s-1)}, h_0(t_i)^{(s-1)})$$

5. Terminate the iteration when $max|\boldsymbol{\zeta}^{(s)} - \boldsymbol{\zeta}^{(s-1)}|$ are small enough. Otherwise, let $s = s + 1$ and go back to step 2.

Before updating parameters in each step, the corresponding $\tilde{Q}_{\lambda,\omega}$ function is approximated by Gaussian quadrature in the E-step.

3.2.5 Tuning Parameter Selection and Two-stage Estimation

Similar to section (2.2.4), I propose to use the BIC-type criterion to determine the values of tuning parameters, where

$$BIC_{\boldsymbol{\lambda}} = -2l_o(\hat{\boldsymbol{\zeta}}) + \log(n) \times df_{\boldsymbol{\lambda}}, \tag{3.17}$$

In (3.17), $\hat{\boldsymbol{\zeta}}$ are the estimators obtained from penalized likelihood defined in Equation (3.11) under the given $\boldsymbol{\lambda}$, and $l_o(\hat{\boldsymbol{\zeta}})$ is the value of the observed likelihood $l_o(\boldsymbol{\zeta})$ evaluated at the estimates $\hat{\boldsymbol{\zeta}}$. The tuning parameters are chosen to minimize the $BIC_{\boldsymbol{\lambda}}$ criterion. The total sample size n is used. I take d, the total number of non-zero estimates of $\hat{\zeta}$ as the degree of freedom $df_{\boldsymbol{\lambda}}$.

To reduce the estimation bias, I propose a two-stage process. In the first stage, I focus on using the penalized likelihood method to perform structural discovery to select the model that minimizes the BIC value. In the second stage, to reduce the estimation bias,

I re-estimate parameters using SSANOVA type penalized likelihood with selected model structure from the first stage.

3.3 Simulation Study

3.3.1 Data Generation

I conducted a simulation study to evaluate the performance of the proposed estimation procedure. I generated the longitudinal outcome Y_{ij} from the following model:

$$\begin{aligned} Y_{ij} =& 2 + 2x_{1ij,1} + sin(x_{1ij,2}) + \frac{(\frac{x_{1ij,3}}{\pi})^{2-1}(1-\frac{x_{1ij,3}}{\pi})^{5-1}}{Beta(2,5)} + \\ & 0x_{1ij,4} + 0x_{1ij,5} + 0x_{1ij,6} + 0x_{1ij,7} + 0x_{1ij,8} + b_{0i} + \epsilon_{ij}, \end{aligned} \tag{3.18}$$

in which, $x_{1ij,1}$, $x_{1ij,2}$, and $x_{1ij,3}$ have the linear, nonlinear and partially linear effects. The failure time was generated from the proportional hazard:

$$\begin{aligned} \lambda_i(t) = \lambda_0(t) \exp\big\{ & 3x_{2i,1} + \frac{(\frac{x_{2i,2}}{\pi})^{5-1}(1-\frac{x_{2i,2}}{\pi})^{2-1}}{Beta(5,2)} + sin(x_{2i,3}) + \\ & 0x_{2i,4} + 0x_{2i,5} + 0x_{2i,6} + 0x_{2i,7} + 0x_{2i,8} + b_{0i}\big\}, \end{aligned} \tag{3.19}$$

in which, $x_{2i,1}$, $x_{2i,2}$, and $x_{2i,3}$ have the linear, partially linear and nonlinear effects, for $i = 1, \ldots, 500, j = 1, \ldots, 5$, where $\lambda_0(t) = \alpha\lambda t^{\alpha-1}$ with $\alpha = 2$, and $\lambda = \exp(1) = 2.718$.

Random intercept b_{0i} was independently generated from $N(0, 1)$. Covariates $x_{1ij,h}$, and $x_{2i,h}, h = 1, \ldots, 8$ were generated from Uniform(0, π) distributions and the measurement error $\epsilon_{ij} \sim i.i.d.N(0, 1)$. Censoring time was independently generated from an exponential distribution to achieve 30% censoring percentage. I generated three different correlations among the independent variables. Let $\rho_1 = corr(x_{1ij,h}, x_{1ij,h'})$ and $\rho_2 = corr(x_{2i,m}, x_{2i,m'})$ denote the correlations between the covariates. In Scenario 1, $\rho_1 = \rho_2 = 0$; in Scenario 2, $\rho_1 = \rho_2 = 0.3$; in Scenario 3, $\rho_1 = \rho_2 = 0.6$.

For each scenario, I generated 100 data sets and applied the proposed method to discover the functional forms of independent variables in the first stage and fitted the second-stage model according to the discovered model structure. The tuning parameters $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ were determined by minimizing the BIC criterion, as defined in (3.17). I also fitted the model using SSANOVA type penalized likelihood without performing structural discovery for comparison.

3.3.2 Simulation Results

I present the structural discovery results of the three scenarios in Table 3.1. In Scenario 1, for the longitudinal component, the discovery accuracy for nonlinear and partially linear effects are 100%. The accuracy of linear effect discovery is 98%. For the five noise independent variables with no effects, four of them have the discovery accuracies of over 98% and the remaining one has the accuracy of 93%. The structural discovery accuracies of independent variables for longitudinal component of Scenario 2 are similar to those of Scenario 1, while Scenario 3 has slightly reduced accuracies. For the survival component, the discovery accuracies in Scenarios 1 (>96%) and 2 (>91%) are excellent for all independent variables. The accuracies of Scenario 3 are slightly lower, and the accuracy is 81% for the partially linear effect.

To evaluate whether the method could improve the estimation of the functional forms, I compare our first- and second-stage models with the SSANOVA model without any structural discovery. For each of the 100 simulated data, I calculate the integrated squared error (ISE) for each independent variable as $ISE_{1h} = E_{X_{1h}}\{f_h(X_{1h}) - \hat{f}_h(X_{1h})\}^2$ and $ISE_{2h} = E_{X_{2h}}\{g_h(X_{2h}) - \hat{g}_h(X_{2h})\}^2$ via a Monte Carlo integration on X_{1h} and X_{2h} for longitudinal and survival components, respectively. X_{1h} and X_{2h} are Uniform(0,π) random variables generated for the simulation study. I then calculate the average ISEs over the 100 real-

izations as $AISE_{1h}$ and $AISE_{2h}$ for each of the three models: the SSANOVA, first-stage and second-stage models. The total AISEs (TAISE) are defined as $TAISE_1 = \sum_{h=1}^{p} AISE_{1h}$ and $TAISE_2 = \sum_{h=1}^{p} AISE_{2h}$. The estimation results are summarized in Table 3.2. For longitudinal component, the TAISEs of first-stage model are much smaller than those of SSANOVA model and the second-stage model could further improve the estimation with even reduced TAISEs. When the pairwise correlation of the independent variables increases, the TAISEs increase, but both the first-stage and second-stage model still perform much better than the SSANOVA model. For survival component, the TAISEs of first-stage model is larger than that of SSANOVA model, which is probably due to the increased bias caused by introducing penalty terms for structural discovery to the likelihood. The bias could be greatly reduced by the second-stage model. The TAISEs of the second-stage model are the smallest among the three models. The pairwise correlation of independent variables seems to have more influence on the survival component. The average curve estimate of functional form for each independent variable over the 100 simulated data sets are shown in Figures 3.1, 3.2, 3.3, 3.4, 3.5, and 3.6. As shown in the figures, for the independent variables with nonzero effects, the SSANOVA and second-stage model estimates are very close to the true curves. However, for the independent variables with no effects, the SSANOVA model has much more pronounced bias than the second-stage model.

Computation time of the proposed method is reasonable given the complexity of the joint models. I used 3 quadrature points for Gaussian quadrature integration. With 3 quadrature points, each data set took approximately 90 minutes to complete the first-stage structural discovery under one tuning parameter, and it took another 1 minute in the second-stage estimation under one tuning parameter. The computing time is estimated based on a single CPU (Intel(R) Xeon(R) CPU E7- 4830 @ 2.13GHz) and 4 GB memory in the Unix system. The total computing time depended on the number of tuning parameters. Other

factors, such as the number of random effects could also influence the approximation accuracy of Gaussian quadrature and the computing time. The reduction of computation time from first-stage to second-stage model also reflects the power of our method to construct a parsimonious model by clearly identifying the functional form of each independent variable.

In summary, the structural discovery accuracy is close to 100% for the longitudinal component, and it is robust to the existence of noise variables even when the pairwise correlation of independent variables is strong. For the survival component, the structural discovery accuracy is nearly perfect when pairwise correlation is low or moderate. When the correlation is high, the structural discovery accuracy decreases slightly for the partially linear effect. The structural discovery accuracy in the survival component is also robust to the existence of noise variables. The estimation of the functional forms could be improved though the two-stage procedure, and the improvement is more significant when the pairwise correlation is stronger. In summary, I contend that the proposed structural discovery method works well even under strong correlations of independent variables. The two-stage procedure ensures good structural discovery performance in the first stage and reduced estimation error in the second stage.

Table 3.1: Structural discovery accuracy in longitudinal and survival components for Scenarios 1 to 3

	Discovery Accuracy (%) for Longitudinal component							
Scenarios	$2X_{1,1}$	$sin(X_{1,2})$	$\frac{(\frac{X_{1,3}}{\pi})(1-\frac{X_{1,3}}{\pi})^4}{Beta(2,5)}$	0	0	0	0	0
	L[a]	NL[a]	PL[a]	Noise				
1. $\rho = 0$	98	100	100	98	98	99	93	100
2. $\rho = 0.3$	95	100	100	98	99	96	99	98
3. $\rho = 0.6$	92	100	100	92	94	94	93	92
	Discovery Accuracy (%) for Survival component							
Scenarios	$2X_{2,1}$	$\frac{(\frac{X_{2,3}}{\pi})^4(1-\frac{X_{2,3}}{\pi})}{Beta(5,2)}$	$sin(X_{2,2})$	0	0	0	0	0
	L[a]	PL[a]	NL[a]	Noise				
1. $\rho = 0$	97	100	100	96	98	97	96	100
2. $\rho = 0.3$	98	91	96	99	99	98	100	99
3. $\rho = 0.6$	92	81	97	96	96	98	96	97

[a] L: linear effect; NL: nonlinear effect; PL: partial linear effect

Table 3.2: TAISE of longitudinal and survival components for Scenarios 1 to 3

	Longitudinal component		
Scenarios	SSANOVA	1*st* Stage	2*nd* Stage
1. $\rho = 0$	0.17544	0.14995	0.13679
2. $\rho = 0.3$	0.17684	0.13609	0.13186
3. $\rho = 0.6$	0.21076	0.14486	0.14976
	Survival component		
Scenarios	SSANOVA	1*st* Stage	2*nd* Stage
$\rho = 0$	0.45742	0.58991	0.44554
$\rho = 0.3$	0.47172	0.62455	0.46467
$\rho = 0.6$	0.61915	0.67122	0.55319

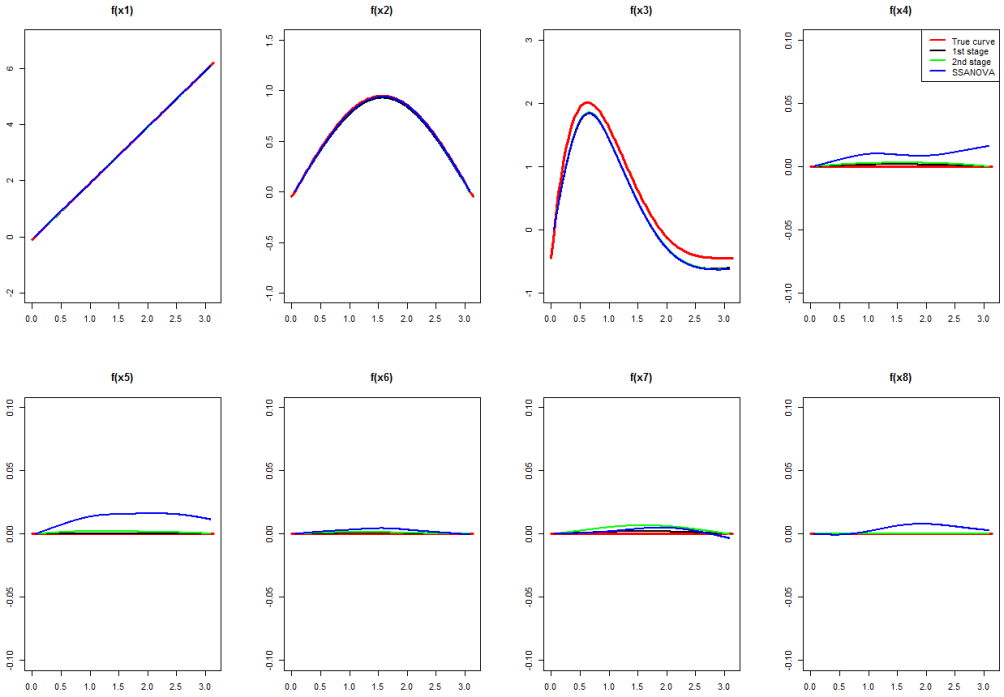

Figure 3.1: Curve estimates in the longitudinal component for Scenario 1.

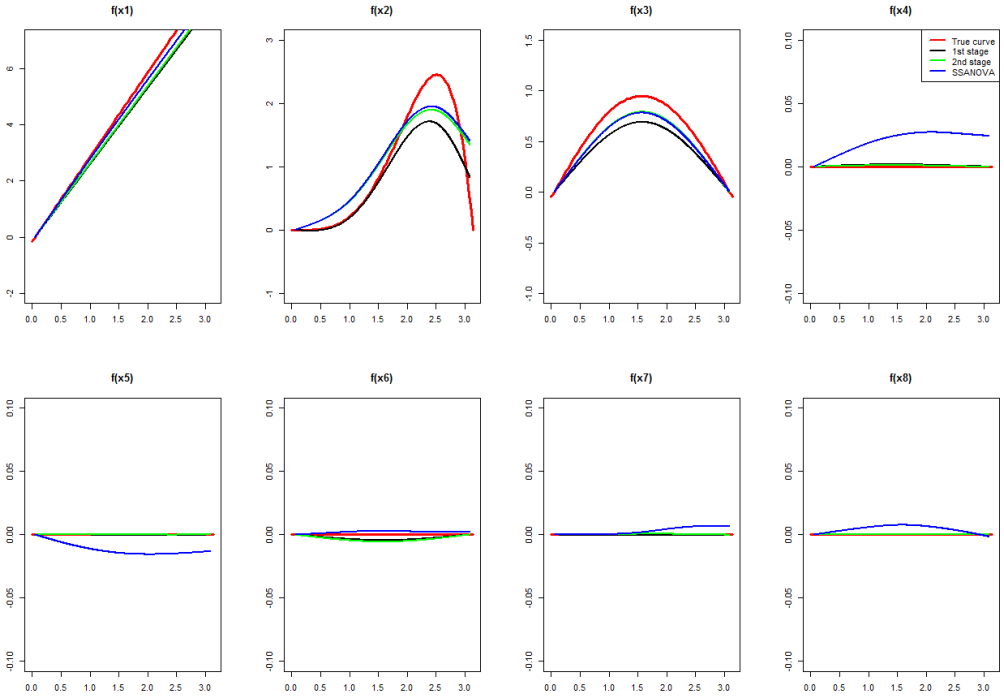

Figure 3.2: Curve estimates in the survival component for Scenario 1.

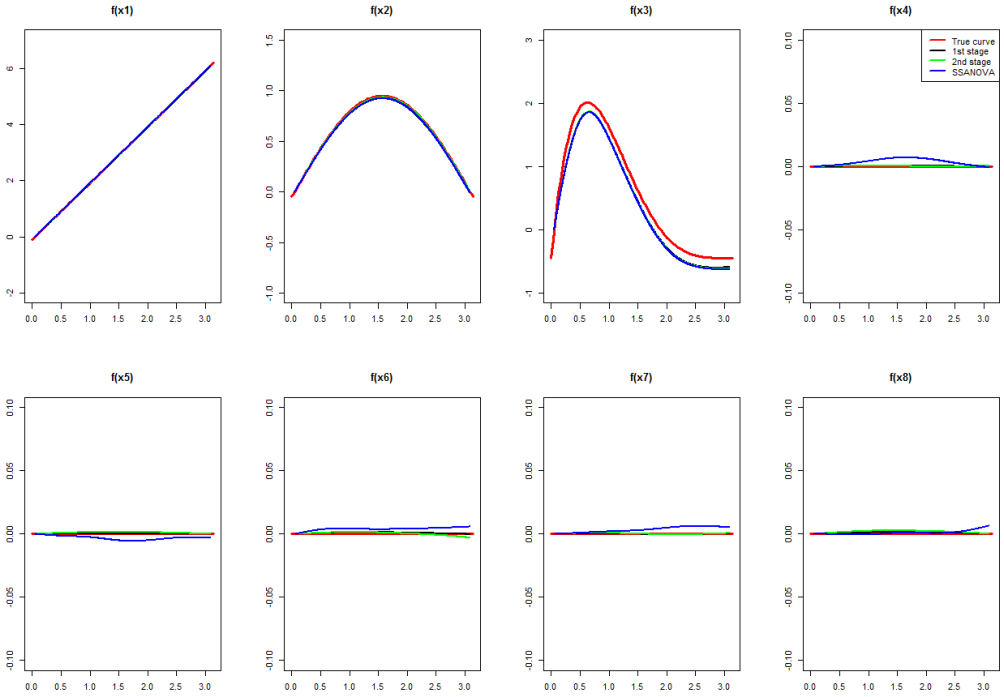

Figure 3.3: Curve estimates in the longitudinal component for Scenario 2.

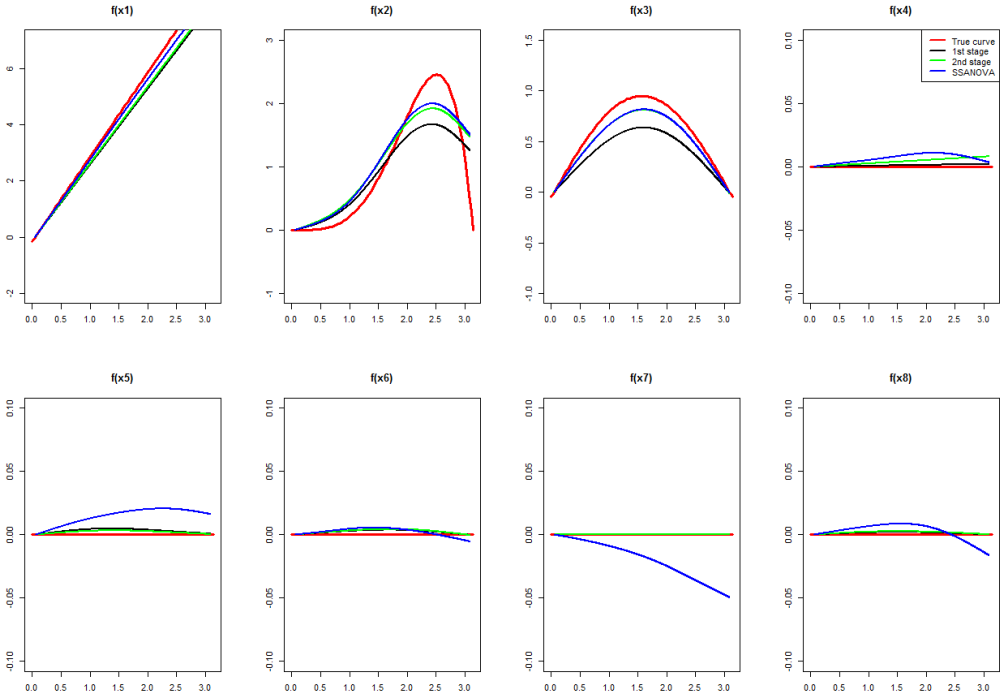

Figure 3.4: Curve estimates in the survival component for Scenario 2.

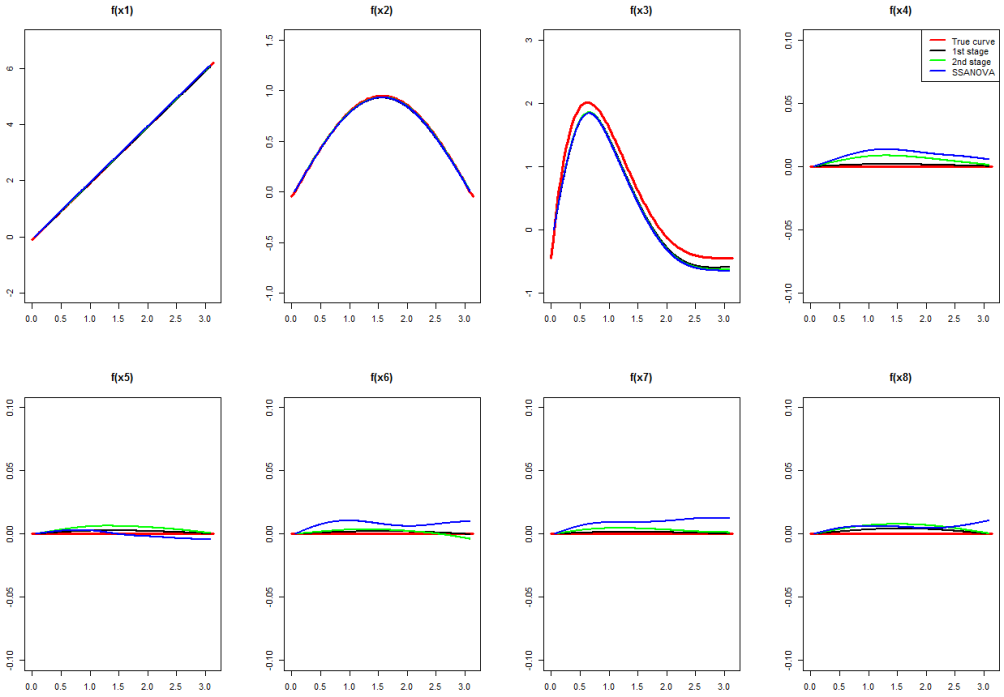

Figure 3.5: Curve estimates in the longitudinal component for Scenario 3.

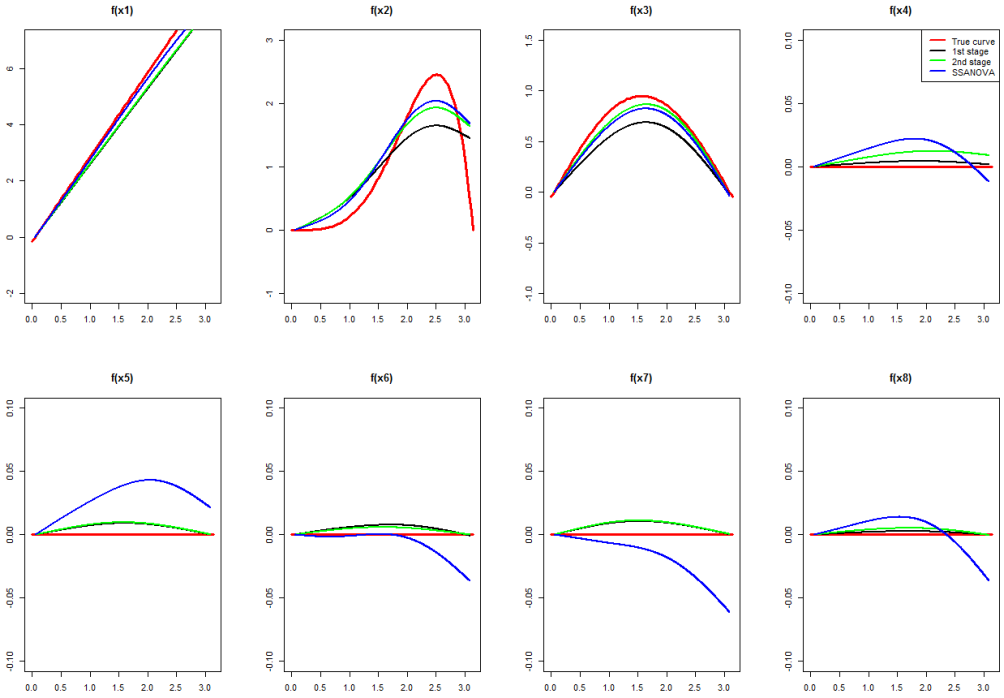

Figure 3.6: Curve estimates in the survival component for Scenario 3.

3.4 Discussion

The proposed structural discovery method provides a practically useful tool for joint models. It helps analysts to achieve parsimonious models without sacrificing the accommodation of nonlinear effects. The method uses penalized likelihood method for sparse computation for the linear and nonlinear elements of cubic B-splines, respectively. Linear and nonlinear elements are separated using the spectral decomposition, and through this, I bridge the selection of linear and nonlinear elements with the simultaneous selection of fixed and random effect in mixed model in the joint model setting (He et al., 2014). The proposed method performs as expected in identifying linear, nonlinear effects, or their combination, the partially linear effect with good accuracy and very little bias. Computationally, the estimation procedure is robust as the algorithm converges 100% in all simulation settings at the threshold of 10^{-7}.

The proposed method has the potential to be adopted or extended in several different directions. First, the method is applicable for recognizing functional forms for additive models. Although in most applications, additive model assumption is sufficient, researchers may be interested in the joint influence of independent variables on the response variables. For example, within the current additive model setting, we are unable to model the interaction between two independent variables. How to jointly determine the functional forms of independent variables would be of interest and practically important for the future work. Another natural extension is for single index models. The current methodology can be extended to recognize the functional forms of the variables involved in the linear predictors, which gives analysts increased modeling flexibility.

In summary, I have shown that representation of cubic B-spline by mixed-effect model and penalized likelihood method could be used for structural discovery in joint model settings. Our research has demonstrated, through a simulation study, that the proposed

method provides a useful structural discovery tool and the method is easy to implement and efficient in computation.

Chapter 4

Selection of Time-Varying Coefficients

4.1 Introduction

Linear regression is one of the most frequently used approaches in scientific research. Linear models have well-established properties. However, the assumed linear relationship is not always realistic in practice. When the linear model assumption is violated, the linear model will be at risk of being mis-specified, which could result in questionable results. Nonparametric models, which require fewer model assumptions, offer much greater modeling flexibility. But when the dimension of independent variable is high, nonparametric models could be subject to increased modeling complexity, which makes the model harder to fit and interpret. An alternative approach is to loosen the restrictions on the fixed linear effect and let the regression coefficient vary as a nonlinear function. Such a model can be viewed as an extension of the traditional linear regression model. It is often referred to as the varying coefficient model (Hastie and Tibshirani, 1993). Varying coefficient model depicts the effect of an independent variable as a function of another independent variable instead of being a constant. The appeals of varying coefficient models stem not only from mathematical elegance, but also from the practical needs. In many situations, it might not be reasonable to assume constant effects of independent variables, and this is typically the case for more complex biological systems. Varying coefficient models greatly enhance the modeling flexibility and they have been widely used for discovering nonlinear effects that would have been missed by traditional parametric models. For example, the effect of sodium-retaining hormone aldosterone on blood pressure may be dependent on the prevailing levels of extracellular fluid volume, as reflected by plasma renin activity. Varying

coefficient model provides a flexible modeling framework to accommodate such interacting influences (Tu et al., 2014). But more often, by allowing the effects of certain independent variables on the outcome to vary over time, this new class of models provides the necessary flexibility to depict time-changing effects of independent variables. Similarly, such a need may extend to survival analysis, for needs to model the time-dependent effect of an independent variable. For example, in an analysis of sexually transmitted infections, Yu et al. (2012) showed the effect of number of partners on infection acquisition tended to be age-dependent. In a childhood asthma study, the effect of airway reactivity measurement on the risk of wheezing is known to change over time due to child growth (Yu et al., 2013).

In practice, if a study collects a large number of independent variables, it is not always feasible to assume all of them have the time-varying effects. Modeling all independent variables with time-varying coefficients could significantly decrease the modeling efficiency and interpretability. If an independent variable effect is approximately constant over time, a time-invariant coefficient is clearly preferable. In addition, I also want to exclude from the model variables that have no effects on the outcome. In joint models of longitudinal and survival outcomes, both model components may have independent variables that interact with time. A statistical method that consistently selects the important variables and correctly distinguishes their temporal effects as time invariant or time varying would be helpful in the construction of joint models. To the best of our knowledge, no such work has been done for joint models.

Literature of model selection for time-varying coefficients is limited. For longitudinal data analysis, Leng (2009) proposed a method to select between time-invariant and time-varying coefficients, but the method does not exclude zero coefficients. Wang et al. (2008) proposed a penalized likelihood method with smoothly clipped absolute deviation (SCAD, Fan and Li 2001) penalty on the expanded nonparametric basis functions of coefficients.

For the survival analysis, Leng and Helen Zhang (2006) extended the component selection and applied smoothing operator proposed by Lin and Zhang (2006) to the Cox model with varying coefficients. More recently, Yan and Huang (2012) proposed to use adaptive LASSO to select time-invariant and time-varying coefficients as well as exclude the zero coefficients in the Cox model. In this research, I develop a penalized likelihood method to simultaneously distinguish the time-invariant coefficients from time-varying coefficients in both the longitudinal and survival components in joint models. I propose to use B-spline to model the unknown temporal effects of independent variables and decompose the B-spline into the time-invariant and time-varying parts, then use the variable selection tools to select the two parts separately to distinguish the time-invariant coefficients from time-varying coefficients. The selection process is similar to mixed-effect selection, which has been described for joint models in Chapter 2.

4.2 Method

4.2.1 Model Formulation

In a longitudinal study, one has a survival outcome (t_{si}, δ_i), and repeated measurements of a continuous outcome $\boldsymbol{y}_i$ for $i = 1, \cdots, n$ subjects, measured at a series of times $\boldsymbol{t}_i$. Here t_{si} is the observed event time subjected to right censoring, and δ_i is a failure indicator with $\delta_i = 1$ indicating the occurrence of an event of interest, and $\delta_i = 0$ indicating censoring, whereas $\boldsymbol{y}_i = \{y_{i1} \dots y_{in_i}\}$ is an $n_i \times 1$ vector of the n_i repeated measurements and $\boldsymbol{t}_i = (t_{i1}, t_{i1}, \dots, t_{in_i})$ is a vector of corresponding measuring times. For the longitudinal component, I denote the p independent variables with unknown temporal effects as $\boldsymbol{x_{1ij}} = (x_{1ij,1}, x_{1ij,2}, \dots, x_{1ij,p})$, which are measured at time t_{ij}, for $j = 1, \dots, n_i$. For the survival component, I denote the p independent variables with unknown temporal effects measured at event time or censoring time t_{si} as $\boldsymbol{x_{2i}} = (x_{2i,1}, x_{2i,2}, \dots, x_{2i,p})$. Without loss

of generality, I herein consider a case where the longitudinal and survival components share the same set of independent variables. This model formulation could easily be generalized to situations where the two components have different sets of independent variables. I denote the random effect covariate vectors for longitudinal and survival components as $\boldsymbol{r_{1ij}} \in \mathbb{R}_1^q$ and $\boldsymbol{r_{2i}} \in \mathbb{R}_1^q$, respectively. I would then construct the time-varying coefficient joint model as follows.

I present the time-varying coefficient linear mixed-effects model for longitudinal component as:

$$y_{ij} = \beta_0 + \sum_{k=1}^{p} \beta_{1k}(t_{ij}) x_{1ij,k} + \boldsymbol{r_{1ij}^T b_i} + \varepsilon_{ij}, \tag{4.1}$$

and the hazard function with time-varying coefficients for survival component as:

$$h(t_{si}) = h_0(t_{si}) \exp(\sum_{k=1}^{p} \beta_{2k}(t_{si}) x_{2i,k} + \boldsymbol{r_{2i}^T b_i}) \tag{4.2}$$

where β_0 is the intercept; $\beta_{1k}(t_{ij})$s and $\beta_{2k}(t_{si})$s are the regression coefficients for unknown temporal effects in the two components; $\boldsymbol{b_i} = (b_1, \cdots, b_q) \in \mathbb{R}_1^q$ is a $q-$dimensional random effect vector following a multivariate normal distribution $MVN(\boldsymbol{0, D(\phi)})$, where $\boldsymbol{D(\phi)}$ is the random effect covariance matrix; when $\boldsymbol{b_i}$ is given, I assume that longitudinal outcome Y_i and survival outcome T_{si} are independent; $\varepsilon_{ij} \sim N(0, \sigma^2)$ is the measurement error, which is assumed to be independently and identically distributed, and $h_0(t_{si})$ is the baseline hazard.

4.2.2 Representing the Model by Decomposed B-spline

To determine the temporal effects, one needs to identify the forms of $\beta_{1k}(t)$ and $\beta_{2k}(t)$. To perform this, I firstly model $\beta_{1k}(t)$ and $\beta_{2k}(t)$ by B-spline and then decompose the B-spline into two parts: one is used to explain the time-invariant effect, and the other is used to

explain the time-varying effect. The decomposition would allow us to use variable selection method to select the two parts separately, and thus discriminating the two effects.

To illustrate the decomposition of the B-splines, I use $\beta_{1k}(t)$ as an example and note that $\beta_{2k}(t)$ could be decomposed and represented with the same approach. By following Yan and Huang (2012), I assume $\beta_{1k}(t) = \boldsymbol{B_1(t)\theta_{1k}}$, $k = 1 \ldots, p$, with $\boldsymbol{B_1} = (1, B_{11}(t), \ldots, B_{1q-1}(t))$. $\tilde{\boldsymbol{B}}_{\boldsymbol{1}} = (B_{11}(t), \ldots, B_{1q-1}(t))$ $(q > 1)$ is a set of B-spline basis of $q-1$ degrees of freedom on a predetermined time interval and knots without intercept, and I could rewrite $\boldsymbol{B_1} = (1, \tilde{\boldsymbol{B}}_{\boldsymbol{1}})$. The corresponding regression coefficients $\boldsymbol{\theta_{1k}}$ are decomposed as $\boldsymbol{\theta_{1k}} = (\theta_{1k,1}, \boldsymbol{\theta_{1k,-1}})$. I then represent the coefficient $\beta_{1k}(t)$ as $\beta_{1k}(t) = \theta_{1k,1} + \tilde{\boldsymbol{B}}_{\boldsymbol{1}}\boldsymbol{\theta_{1k,-1}}$. Through this decomposition, I construct the nonparametric function of $\beta_{1k}(t)$ in such a way that $\theta_{1k,1}$ is associated with the intercept in $\boldsymbol{B_1}$ representing an overall time-invariant effect, whereas $\boldsymbol{\theta_{1k,-1}}$ represents the time-varying effect relative to the intercept. As a result, I could represent the linear mixed-effects model for longitudinal component defined in Equation (4.1) by decomposed B-splines as:

$$\begin{aligned} y_{ij} &= \beta_0 + \sum_{k=1}^{p} \boldsymbol{B_1(t_{ij})\theta_{1k}} \cdot x_{1ij,k} + \boldsymbol{r_{1ij}^T b_i} + \varepsilon_{ij} \\ &= \beta_0 + \sum_{k=1}^{p} [\theta_{1k,1} + \tilde{\boldsymbol{B}}_{\boldsymbol{1}}\boldsymbol{(t_{ij})\theta_{1k,-1}}] \cdot x_{1ij,k} + \boldsymbol{r_{1ij}^T b_i} + \varepsilon_{ij} \end{aligned} \tag{4.3}$$

Similarly, for the survival component, I assume $\beta_{2k}(t) = \boldsymbol{B_2(t)\theta_{2k}}$, $k = 1 \ldots, p$, where $\boldsymbol{B_2} = (1, B_{21}(t), \ldots, B_{2q-1}(t))$, and $\tilde{\boldsymbol{B}}_{\boldsymbol{2}} = (B_{21}(t), \ldots, B_{2q-1}(t))$ $(q > 1)$ is a set of B-spline basis of $q-1$ degrees of freedom on a predetermined time interval without intercept. Regression coefficients $\boldsymbol{\theta_{2k}}$ are decomposed as $\boldsymbol{\theta_{2k}} = (\theta_{2k,1}, \boldsymbol{\theta_{2k,-1}})$ and $\beta_{2k}(t)$ is represented as $\beta_{2k}(t) = \theta_{2k,1} + \tilde{\boldsymbol{B}}_{\boldsymbol{2}}\boldsymbol{\theta_{2k,-1}}$. I could represent the model for survival component defined

in Equation (4.2) by decomposed B-splines as:

$$\begin{aligned} h(t_{si}) &= h_0(t_{si}) \exp\{\sum_{k=1}^{p} \boldsymbol{B}_2(t_{si})\boldsymbol{\theta}_{2k} \cdot x_{2i,k} + \boldsymbol{r}_{\boldsymbol{2i}}^{\boldsymbol{T}}\boldsymbol{b}_{\boldsymbol{i}}\} \\ &= h_0(t_{si}) \exp\{\sum_{k=1}^{p} [\theta_{2k,1} + \tilde{\boldsymbol{B}}_{\boldsymbol{2}}(\boldsymbol{t}_{\boldsymbol{si}})\boldsymbol{\theta}_{\boldsymbol{2k,-1}}] \cdot x_{2i,k} + \boldsymbol{r}_{\boldsymbol{2i}}^{\boldsymbol{T}}\boldsymbol{b}_{\boldsymbol{i}}\} \end{aligned} \tag{4.4}$$

Let $\boldsymbol{\eta} = (\beta_0, \boldsymbol{\theta}_{\boldsymbol{1k}}, \boldsymbol{\theta}_{\boldsymbol{2k}}, \phi, \sigma, h_0(t_{si}))$ denote all the unknown parameters. The marginal likelihood of $\boldsymbol{\eta}$ could be obtained as:

$$L_o(\boldsymbol{\eta}) = \prod_{i=1}^{n} \int \prod_{j=1}^{n_i} \Big\{ [f_y(t_{si})|\boldsymbol{b}_i, \boldsymbol{x}_{\boldsymbol{2i}}, \boldsymbol{r}_{\boldsymbol{2i}}, \delta_i, \boldsymbol{\eta}][f_s(t_{si})|\boldsymbol{b}_i, \boldsymbol{x}_{\boldsymbol{2i}}, \boldsymbol{r}_{\boldsymbol{2i}}, \delta_i, \boldsymbol{\eta}] \Big\} f_b(\boldsymbol{b}_{\boldsymbol{i}}) d\boldsymbol{b}_{\boldsymbol{i}}, \tag{4.5}$$

and the log-marginal likelihood is $l_o(\boldsymbol{\eta}) = \log L_o(\boldsymbol{\eta})$, where $f_b(\cdot)$ is a $q-$variate normal density function for $\boldsymbol{b}_{\boldsymbol{i}}$, $f_s(\cdot)$ is the likelihood of survival component parameters conditional on $\boldsymbol{b}_{\boldsymbol{i}}$, and $f_y(\cdot)$ is the conditional distribution of repeated measurements when $\boldsymbol{b}_{\boldsymbol{i}}$ is given. The hazard $h_0(t_{si})$ could be modeled nonparametrically or parametrically as described in Chapter 3.

4.2.3 Selection of Time-Varying Coefficients by Penalized Likelihood

To estimate the nonparametric functions in Equation (4.3) and (4.4), a frequently used method is the ridge type penalized likelihood with quadratic penalty terms:

$$pl_{Ridge}(\boldsymbol{\eta}) = \frac{1}{n} l_o(\boldsymbol{\eta}) - \lambda_1 \sum_{k=1}^{p} ||\boldsymbol{\theta}_{\boldsymbol{1k,-1}}||^2 - \lambda_2 \sum_{k=1}^{p} ||\boldsymbol{\theta}_{\boldsymbol{2k,-1}}||^2 \tag{4.6}$$

However, the penalty terms in (4.6) do not possess the sparsity property and could not perform selection of the time-varying effects. The penalized likelihood (4.6) also lacks the penalty terms for selection of time-invariant coefficients. To perform selection of both the

time-invariant and time-varying coefficients, I propose the following penalized likelihood:

$$pl_o(\boldsymbol{\eta}) = \frac{1}{n} l_o(\boldsymbol{\eta}) - p(\boldsymbol{\theta}) \tag{4.7}$$

where $p(\boldsymbol{\theta}) = \lambda_1 \sum_{k=1}^{p} \kappa_1(\theta_{1k,1}) + \lambda_2 \sum_{k=1}^{p} \kappa_2(\theta_{2k,1}) + \lambda_3 \sum_{k=1}^{p} \kappa_3(\boldsymbol{\theta_{1k,-1}})$
$+\lambda_4 \sum_{k=1}^{p} \kappa_4(\boldsymbol{\theta_{2k,-1}})$. The penalty functions $\kappa_1(\theta_{1k,1})$ and $\kappa_2(\theta_{2k,1})$ control the sparsity of estimates of $\theta_{1k,1}$ and $\theta_{2k,1}$ so that the time-invariant coefficients are selected, where λ_1 and λ_2 are the associated positive tuning parameters. The penalty terms $\kappa_3(\boldsymbol{\theta_{1k,-1}})$ and $\kappa_4(\boldsymbol{\theta_{2k,-1}})$ control the sparsity of estimates of $\boldsymbol{\theta_{1k,-1}}$ and $\boldsymbol{\theta_{2k,-1}}$ to select the time-varying coefficients, where λ_3 and λ_4 are the associated positive tuning parameters. The penalty functions $\kappa_1(\cdot)$, $\kappa_2(\cdot)$, $\kappa_3(\cdot)$, $\kappa_4(\cdot)$ could be the adaptive LASSO, or the smoothly clipped absolute deviation (SCAD).

For the selection of time-invariant coefficients, I define the adaptive LASSO penalty as $\kappa_1(\theta_{1k,1}) = \omega_{1k}|\theta_{1k,1}|$ and $\kappa_2(\theta_{2k,1}) = \omega_{2k}|\theta_{2k,1}|$, where ω_{1k}, ω_{2k} are the corresponding positive weights for penalties $|\theta_{1k,1}|$ and $|\theta_{2k,1}|$. I choose the weights as $\omega_{1k} = 1/|\tilde{\theta}_{1k,1}|, \omega_{2k} = 1/|\tilde{\theta}_{2k,1}|$, where $\tilde{\theta}_{1k,1}$ and $\tilde{\theta}_{2k,1}$ are the optimizers of the ridge type penalized likelihood (4.6). Some of the estimates of $\hat{\theta}_{1k,1}$ and $\hat{\theta}_{2k,1}$ for penalized likelihood (4.7) will be zero since $|\theta_{1k,1}|$ and $|\theta_{2k,1}|$ are singular when $|\theta_{1k,1}| = 0$ and $|\theta_{2k,1}| = 0$.

For the selection of time-varying coefficients, I noted that the time-varying effect could be excluded if and only if $\boldsymbol{\theta_{1k,-1}} = 0$ and $\boldsymbol{\theta_{2k,-1}} = 0$, and I propose to select $\boldsymbol{\theta_{1k,-1}}$ and $\boldsymbol{\theta_{2k,-1}}$ in a group manner. I first summarize the penalty terms using L_2-norm: $||\boldsymbol{\theta_{1k,-1}}|| = (\boldsymbol{\theta_{1k,-1}^T \theta_{1k,-1}})^{1/2}$ and $||\boldsymbol{\theta_{2k,-1}}|| = (\boldsymbol{\theta_{2k,-1}^T \theta_{2k,-1}})^{1/2}$. Following Yuan and Lin (2006), the adaptive LASSO penalties for time-varying coefficients are defined as: $\kappa_3(\boldsymbol{\theta_{1k,-1}}) = \omega_{3k}||\boldsymbol{\theta_{1k,-1}}||$ and $\kappa_4(\boldsymbol{\theta_{2k,-1}}) = \omega_{4k}||\boldsymbol{\theta_{2k,-1}}||$. I choose the weights as $\omega_{3k} = p_{1k}/||\tilde{\boldsymbol{\theta}}_{\boldsymbol{1k,-1}}||$ and $\omega_{4k} = p_{2k}/||\tilde{\boldsymbol{\theta}}_{\boldsymbol{2k,-1}}||$, where $\tilde{\boldsymbol{\theta}}_{\boldsymbol{1k,-1}}$ and $\tilde{\boldsymbol{\theta}}_{\boldsymbol{2k,-1}}$ are the optimizers of the ridge type penalized likelihood (4.6). p_{1k} and p_{2k} are the sizes of $\tilde{\boldsymbol{\theta}}_{\boldsymbol{1k,-1}}$ and $\tilde{\boldsymbol{\theta}}_{\boldsymbol{2k,-1}}$, respectively. I

use adaptive LASSO penalty in the simulation study. Penalized likelihood with the SCAD penalty terms could be constructed in a similar way by substituting the penalty terms in (4.7) with penalties on $(|\theta_{1k,1}|, |\theta_{2k,1}|, ||\boldsymbol{\theta_{1k,-1}}||, ||\boldsymbol{\theta_{2k,-1}}||)$ using SCAD. The estimator of $\boldsymbol{\eta} = (\beta_0, \theta_{1k,1}, \theta_{2k,1}, \boldsymbol{\theta_{1k,-1}}, \boldsymbol{\theta_{2k,-1}}, \boldsymbol{\phi}, \sigma, h_0(t_{si}))$ can be obtained by maximizing Equation (4.7).

4.2.4 Optimization of the Penalized Likelihood

To maximize the penalized likelihood (4.7), I firstly construct the marginal likelihood as:

$$L_o(\boldsymbol{\eta}) = \prod_{i=1}^{n} \int \prod_{j=1}^{n_i} \Big\{ [f_y(y_{ij}|\boldsymbol{b}_i, \boldsymbol{x_{1ij}}, \boldsymbol{r_{1ij}}, \boldsymbol{\eta})][h(t_{si}|\boldsymbol{b}_i, \boldsymbol{x_{2i}}, \boldsymbol{r_{2i}}, \boldsymbol{\eta})]^{\delta_i} \\ [S(t_{si}|\boldsymbol{b}_i, \boldsymbol{x_{2i}}, \boldsymbol{r_{2i}}, \boldsymbol{\eta})] \Big\} f_b(\boldsymbol{b_i}|\boldsymbol{\eta}) d\boldsymbol{b_i}, \tag{4.8}$$

where $h(t_{si}|\boldsymbol{b}_i, \boldsymbol{x_{2i}}, \boldsymbol{r_{2i}}, \boldsymbol{\eta})$ is the hazard function. $S(t_{si}|\boldsymbol{b}_i, \boldsymbol{x_{2i}}, \boldsymbol{r_{2i}}, \boldsymbol{\eta})$ is the survival function, which is expressed as:

$$\begin{aligned}
&S(t_{si}|\boldsymbol{b}_i, \boldsymbol{x_{2i}}, \boldsymbol{r_{2i}}, \boldsymbol{\eta}) \\
&= \exp\Big\{ - \int_0^{t_{si}} h(u) du \Big\} \\
&= \exp\Big\{ - \int_0^{t_{si}} h_0(u) \exp\{\sum_{k=1}^{p} \beta_{2k}(u) \cdot x_{2i,k} + \boldsymbol{r_{2i}^T b_i}\} du \Big\} \\
&= \exp\Big\{ - \exp(\boldsymbol{r_{2i}^T b_i}) \int_0^{t_{si}} h_0(u) \exp\{\sum_{k=1}^{p} \boldsymbol{B}_2(u)\boldsymbol{\theta}_{2k} \cdot x_{2i,k}\} du \Big\} \\
&= \exp\Big\{ - \exp(\boldsymbol{r_{2i}^T b_i}) \int_0^{t_{si}} h_0(u) \exp\{\sum_{k=1}^{p} [\theta_{2k,1} + \boldsymbol{\tilde{B}_2(u)\theta_{2k,-1}}] \cdot x_{2i,k}\} du \Big\} \\
&= \exp\Big\{ - \exp(\boldsymbol{r_{2i}^T b_i} + \sum_{k=1}^{p} \theta_{2k,1} \cdot x_{2i,k}) \int_0^{t_{si}} h_0(u) \exp\{\sum_{k=1}^{p} [\boldsymbol{\tilde{B}_2(u)\theta_{2k,-1}}] \cdot x_{2i,k}\} du \Big\}.
\end{aligned} \tag{4.9}$$

Because both the survival function and the marginal likelihood involve intractable integrals, I propose to use a Gaussian quadrature method to approximate the intractable integrals.

By using Gauss-Legendre quadrature, I could approximate the intractable integral in the survival function (4.9) as

$$\begin{aligned}
&\int_0^{t_{si}} h_0(u)\exp\{\sum_{k=1}^{p}[\tilde{\boldsymbol{B}}_{\boldsymbol{2}}(\boldsymbol{u})\boldsymbol{\theta}_{\boldsymbol{2k,-1}}]\cdot x_{2i,k}\}du \\
&= \frac{t_{si}-0}{2}\int_{-1}^{1} h_0(\frac{t_{si}-0}{2}z+\frac{t_{si}+0}{2})\exp\{\sum_{k=1}^{p}\tilde{\boldsymbol{B}}_2(\frac{t_{si}-0}{2}z+\frac{t_{si}+0}{2})\boldsymbol{\theta}_{2k,-1}\cdot x_{2i,k}\}dz \\
&\approx \frac{t_{si}}{2}\sum_{h=1}^{n_{node}}\omega_h h_0[\frac{t_{si}}{2}(z_h+1)]\exp\{\sum_{k=1}^{p}\tilde{\boldsymbol{B}}_2[\frac{t_{si}}{2}(z_h+1)]\boldsymbol{\theta}_{2k,-1}\cdot x_{2i,k}\},
\end{aligned} \tag{4.10}$$

where z_h and w_h are the corresponding Gauss-Legendre quadrature nodes and weights.

To approximate the marginal likelihood, I use multivariate Gaussian-Hermite quadrature to integrate the random effect $\boldsymbol{b_i}$ out by following (Pinheiro and Bates, 1995). Since $\boldsymbol{b_i}\in\mathbb{R}_1^q$, if I choose k quadrature points, there will be a total of k^q vector nodes of $q\times 1$ dimension. Let $\boldsymbol{b}_l^{'}=(b_{l,1}^{'},b_{l,2}^{'},\cdots,b_{l,q}^{'})$ denote the lth node, w_l denote the corresponding quadrature weight, for $l=1,\cdots,k^q$. Let g_{oi} be all the observed data and let

$$\begin{aligned}
A(g_{oi}|\boldsymbol{b_i},\boldsymbol{\eta}) =&[f_y(y_{ij})|\boldsymbol{b}_i,\boldsymbol{x_{1ij}},\boldsymbol{r_{1ij}},\boldsymbol{\eta}] \\
&[h_s(t_{si})|\boldsymbol{b}_i,\boldsymbol{x_{2i}},\boldsymbol{r_{2i}},\boldsymbol{\eta}]^{\delta_i}[S_s(t_{si})|\boldsymbol{b}_i,\boldsymbol{x_{2i}},\boldsymbol{r_{2i}},\boldsymbol{\eta}],
\end{aligned}$$

then the integral (4.8) can be approximated by:

$$\int A(g_{oi}|\boldsymbol{b_i},\boldsymbol{\eta})f_b(\boldsymbol{b_i}|\boldsymbol{\eta})d\boldsymbol{b_i}\approx\sum_{l=1}^{k^q}w_l A(g_{oi}|\boldsymbol{b_l'},\boldsymbol{\eta})f_b(\boldsymbol{b_l'}|\boldsymbol{\eta}). \tag{4.11}$$

The approximated penalized likelihood is

$$pl(\boldsymbol{\eta})=\sum_{i=1}^{n}\Big\{\log\sum_{l=1}^{k^q}w_l A(g_{oi}|\boldsymbol{b_l'},\boldsymbol{\eta})f_b(\boldsymbol{b_l'}|\boldsymbol{\eta})\Big\}-p(\boldsymbol{\theta}) \tag{4.12}$$

I maximize Equation (4.12) with respect to $(|\theta_{1k,1}|, |\theta_{2k,1}|)$ by applying the LARS/ LASSO algorithm (Efron et al., 2004), and maximize Equation (4.12) with respect to $(||\boldsymbol{\theta_{1k,-1}}||, ||\boldsymbol{\theta_{2k,-1}}||)$ by following Wang et al. (2010) and Chapter 3.

4.2.5 Tuning Parameter Selection and Two-stage Estimation

Similar to Chapters 2 and 3, I propose to use the BIC-type criterion to determine the values of tuning parameters, where

$$BIC_{\boldsymbol{\lambda}} = -2l_o(\hat{\boldsymbol{\eta}}) + \log(n) \times df_{\boldsymbol{\lambda}}, \tag{4.13}$$

In (4.13), $\hat{\boldsymbol{\eta}}$ is a vector of the estimates obtained from penalized likelihood under a given $\boldsymbol{\lambda}$, and $l_o(\hat{\boldsymbol{\eta}})$ is the value of observed likelihood $l_o(\boldsymbol{\eta})$ at the estimated value of $\hat{\boldsymbol{\eta}}$. The solution is chosen to minimize the $BIC_{\boldsymbol{\lambda}}$ criterion. In this BIC-type criterion, the total sample size n is used. I take d, the total number of non-zero estimates of $\hat{\theta}$ as the degree of freedom $df_{\boldsymbol{\lambda}}$.

As in previous chapters, a two-stage process is used to reduce estimation bias. In the first stage, I perform selection of time-invariant and time-varying coefficients using the proposed penalized likelihood (4.7) to select the model that minimizes the BIC value. In the second stage, I refit the model with selected time-invariant and time-varying effects through the ridge type penalized likelihood to reduce the estimation bias.

4.3 Simulation Study

4.3.1 Data Generation

I conducted a simulation study to examine the performance of the proposed method. I generated the longitudinal outcome Y_{ij} from the following model:

$$\begin{aligned} Y_{ij} =& 5 + \beta_{11}(t)x_{1ij,1} + \beta_{12}(t)x_{1ij,2} + \beta_{13}(t)x_{1ij,3} + \\ & \beta_{14}(t)x_{1ij,4} + \beta_{15}(t)x_{1ij,5} + \beta_{16}(t)x_{1ij,6} + \beta_{17}(t)x_{1ij,7} + \beta_{18}(t)x_{1ij,8} + b_{0i} + \epsilon_{ij}, \end{aligned} \tag{4.14}$$

and the failure time from the distribution with the hazard function:

$$\begin{aligned} \lambda_i(t) = \lambda_0(t) \exp\big\{ & \beta_{21}(t)x_{2i,1} + \beta_{22}(t)x_{2i,2} + \beta_{23}(t)x_{2i,3} + \\ & \beta_{24}(t)x_{2i,4} + \beta_{25}(t)x_{2i,5} + \beta_{26}(t)x_{2i,6} + \beta_{27}(t)x_{2i,7} + \beta_{28}(t)x_{2i,8} + b_{0i} \big\}, \end{aligned} \tag{4.15}$$

for $i = 1, \ldots, 500, j = 1, \ldots, 5$, where $\lambda_0(t) = \alpha\lambda t^{\alpha-1}$ with $\alpha = 2$, and $\lambda = \exp(1) = 2.718$. In the longitudinal component, the coefficients $\beta_{11}(t) = 5$, $\beta_{12}(t) = 3.5 + \frac{(\frac{t}{2})^{5-1}(1-\frac{t}{2})^{2-1}}{Beta(5,2)}$, $\beta_{13}(t) = 2.5 + \frac{(\frac{t}{2})^{2-1}(1-\frac{t}{2})^{5-1}}{Beta(2,5)}$, $\beta_{14}(t) = 0$, $\beta_{15}(t) = 0$, $\beta_{16}(t) = 0$, $\beta_{17}(t) = 0$, and $\beta_{18}(t) = 0$. I generated five measurement times t for each subject from Uniform (0.01,2), plus the baseline measurement t=0. The measurement time could be truncated by the survival time. In the survival component, the coefficients are $\beta_{21}(t) = -2$, $\beta_{22}(t) = 0.5 + \frac{(\frac{t}{2})^{5-1}(1-\frac{t}{2})^{2-1}}{Beta(5,2)}$, $\beta_{23}(t) = -2.5 - \frac{(\frac{t}{2})^{2-1}(1-\frac{t}{2})^{5-1}}{Beta(2,5)}$, $\beta_{24}(t) = 0$, $\beta_{25}(t) = 0$, $\beta_{26}(t) = 0$, $\beta_{27}(t) = 0$, and $\beta_{28}(t) = 0$.

Random intercept b_{0i} was independently generated from $N(0, 0.5)$ distribution. Independent variables $x_{1ij,k}$ and $x_{2i,k}, k = 1, \ldots, 8$ were generated from Uniform(0,1) variables; The measurement error $\epsilon_{ij} \sim i.i.d.N(0, 0.5)$. Censoring times were generated from a mixture distribution of a point mass at 2 and Uniform(0,2) to achieve 25% censoring percentage. I generated two different correlations among the independent variables. Let

$\rho_1 = corr(x_{1ij,k}, x_{1ij,k'})$ and $\rho_2 = corr(x_{2i,k'}, x_{2i,k'})$ denote the correlations between the independent variables. In Scenario 1, I set $\rho_1 = \rho_2 = 0$; in Scenario 2, I set $\rho_1 = \rho_2 = 0.3$. In this simulation study, I used the cubic B-spline.

For each scenario, I generated 100 data sets and applied the proposed method to select the temporal effects for the independent variables and fit the second-stage model according to the selected effects. The tuning parameters $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are determined by minimizing the BIC criterion, as defined in (4.13). I also fit the ridge type penalized likelihood without performing model selection for comparison.

4.3.2 Simulation results

For Scenarios 1 and 2, I present the model selection results in Tables 4.1 and 4.2. For the longitudinal outcome, our method could select the nonzero time-invariant and time-varying coefficients perfectly, as well as exclude the zero ones nearly perfectly, even when there is moderate correlation among the independent variables. In the survival component, for time-invariant coefficient selection, our method selects the non-zero ones with 100% accuracy, and selects the zero ones with 95% accuracy for both scenarios. For time-varying coefficient selection, our method selects non-zero ones with 92% accuracy and selects zero ones with 90% accuracy when there is no correlation among the independent variables. When the correlation increases to 0.3, both the accuracies decrease to about 85%.

To evaluate the estimation performance of the proposed method, I calculated the total average integrated squared error (TAISE) for each estimate of the time-varying coefficients over the 100 simulated data, and the approach was the same as that described in Chapter 3. I compare three procedures, the ridge penalized likelihood without performing model selection, first-stage and second-stage models and report the total TAISEs in Table 4.3. The TAISEs of the proposed method in the second stage are much smaller (about 90% reduction)

Table 4.1: Selection frequency of time-invariant coefficients (TIC)

	Selection Frequency (%) for Longitudinal component							
Scenarios	$\beta_{11}(t)$	$\beta_{12}(t)$	$\beta_{13}(t)$	0	0	0	0	0
	Nonzero TIC			Zero TIC				
$\rho = 0$	100	100	100	0	0	0	0	0
$\rho = 0.3$	100	100	100	0	0	0	0	0
	Selection Frequency (%) for Survival component							
Scenarios	$\beta_{21}(t)$	$\beta_{22}(t)$	$\beta_{23}(t)$	0	0	0	0	0
	Nonzero TIC			Zero TIC				
$\rho = 0$	100	100	100	2	4	4	3	2
$\rho = 0.3$	100	100	100	6	5	3	5	2

as compared to the ridge estimates in the longitudinal component. In the survival component, the TAISEs of the second stage are reduced (about 10% reduction) when compared to the ridge estimates. To evaluate the estimation performance of time-invariant coefficients, I present the average of the estimates and the empirical standard deviation over the 100 data sets in Tables 4.4 and 4.5. In the longitudinal component, the ridge estimates have 10% to 20% biases, and these biases are much reduced in the second-stage estimates (less than 5%). In the survival component, the biases of the ridge estimates for time-invariant coefficients in $\beta_{22}(t)$ and $\beta_{23}(t)$ are much larger, and could not be reduced in the second-stage estimates. One possible reason for the suboptimal estimation performance for the intercept in survival component might be due to the lack of sufficient survival outcomes that are close to time zero, which makes it difficult to accurately estimate the time-invariant effect (the intercept) in the presence of time-varying effects.

Table 4.2: Selection frequency of time-varying coefficients (TVC)

	Selection Frequency (%) for Longitudinal component							
Scenarios	$\beta_{11}(t)$	$\beta_{12}(t)$	$\beta_{13}(t)$	$\beta_{14}(t)$	$\beta_{15}(t)$	$\beta_{16}(t)$	$\beta_{17}(t)$	$\beta_{18}(t)$
	Zero TVC	Nonzero TVC		Zero TVC				
$\rho = 0$	0	100	100	0	0	1	0	0
$\rho = 0.3$	0	100	100	0	0	1	2	0
	Selection frequency (%) for Survival component							
Scenarios	$\beta_{21}(t)$	$\beta_{22}(t)$	$\beta_{23}(t)$	$\beta_{24}(t)$	$\beta_{25}(t)$	$\beta_{26}(t)$	$\beta_{27}(t)$	$\beta_{28}(t)$
	Zero TVC	Nonzero TVC		Zero TVC				
$\rho = 0$	1	100	92	9	5	6	10	10
$\rho = 0.3$	8	100	85	14	14	6	13	13

Table 4.3: TAISE in longitudinal and survival components for time-varying coefficients

	Longitudinal component		
Scenarios	SSANOVA	$1st$ Stage	$2nd$ Stage
$\rho = 0$	0.84214	0.04730	0.07720
$\rho = 0.3$	1.10202	0.07870	0.11380
	Survival component		
Scenarios	SSANOVA	$1st$ Stage	$2nd$ Stage
$\rho = 0$	1.80373	1.63000	1.61000
$\rho = 0.3$	1.81513	1.70272	1.64596

Table 4.4: Estimation results of nonzero time-invariant coefficients (TIC)

	Estimation results for Longitudinal component			
Scenarios		$\beta_{11}(t)$	$\beta_{12}(t)$	$\beta_{13}(t)$
	TIC True value	5	3.5	2.5
$\rho = 0$	SSANOVA	4.9012±0.0572	3.7284±0.0557	2.9641±0.0488
	1^{st} stage	5.0015±0.0410	3.5198±0.0796	2.5088±0.0773
	2^{nd} stage	5.0006±0.0447	3.4935±0.0561	2.5887±0.0523
$\rho = 0.3$	SSANOVA	4.8823±0.0639	3.7462±0.0584	3.0423±0.0599
	1^{st} stage	4.9954±0.0445	3.5061±0.0651	2.5006±0.0619
	2^{nd} stage	5.0031±0.0418	3.4810±0.0646	2.6121±0.0628
	Estimation results for Survival component			
Scenarios		$\beta_{21}(t)$	$\beta_{22}(t)$	$\beta_{23}(t)$
	TIC True value	-2	0.5	-2.5
$\rho = 0$	SSANOVA	-2.0054±0.1958	1.6885±0.1890	-3.3324±0.2173
	1^{st} stage	-1.8925±0.1883	1.3669±0.2825	-3.1082±0.2321
	2^{nd} stage	-1.9942±0.1777	1.6642±0.2048	-3.3107±0.2287
$\rho = 0.3$	SSANOVA	-2.0297±0.2059	1.6587±0.1962	-3.3285±0.2236
	1^{st} stage	-1.8818±0.1944	1.3124±0.2813	-3.1063±0.2276
	2^{nd} stage	-1.9955±0.1880	1.6423±0.2051	-3.3034±0.2326

Table 4.5: Estimation results of zero time-invariant coefficients (TIC)

Scenarios		Estimation results for Longitudinal component				
		$\beta_{14}(t)$	$\beta_{15}(t)$	$\beta_{16}(t)$	$\beta_{17}(t)$	$\beta_{18}(t)$
	TIC True value	0	0	0	0	0
$\rho = 0$	SSANOVA	-0.1009±0.0504	-0.1044±0.0478	-0.1026±0.0594	-0.0973±0.0515	-0.0980±0.0458
	1^{st} stage	0±0	0±0	0±0	0±0	0±0
	2^{nd} stage	0±0	0±0	0±0	0±0	0±0
$\rho = 0.3$	SSANOVA	-0.0967±0.0595	-0.1112±0.0613	-0.1185±0.0511	-0.1230±0.0563	-0.1227±0.0592
	1^{st} stage	0±0	0±0	0±0	0±0	0±0
	2^{nd} stage	0±0	0±0	0±0	0±0	0±0

Scenarios		Estimation results for Survival component				
		$\beta_{24}(t)$	$\beta_{25}(t)$	$\beta_{26}(t)$	$\beta_{27}(t)$	$\beta_{28}(t)$
	TIC True value	0	0	0	0	0
$\rho = 0$	SSANOVA	0.0198±0.2071	0.0368±0.2111	-0.0031±0.2001	0.0478±0.2117	0.0273±0.2320
	1^{st} stage	0.0008±0.0367	0.0120±0.0674	-0.0090±0.0552	-0.0030±0.0271	-0.0019±0.0569
	2^{nd} stage	0.0009±0.0676	0.0228±0.1143	-0.0130±0.1056	-0.0043±0.0764	-0.0014±0.0817
$\rho = 0.3$	SSANOVA	0.0002±0.2343	0.0136±0.2276	-0.0313±0.2117	0.0441±0.2379	-0.0039±0.2489
	1^{st} stage	-0.0030±0.0489	0.0106±0.0676	-0.0092±0.06	0.0080±0.0416	-0.0026±0.0269
	2^{nd} stage	-0.0078±0.1176	0.0185±0.1223	-0.0164±0.0976	0.0164±0.1033	-0.0015±0.0658

4.4 Discussion

In this research, I developed a model selection method to simultaneously identify time-invariant and time-varying effects in a joint model setting through the penalized likelihood approach. As indicated by the simulation study, the method has achieved good accuracy in distinguishing the time-invariant effects from time-varying effects for both longitudinal and survival model components. The method provides a tool that is particularly useful for the analysts who are interested in evaluating the temporal effects of independent variables. This approach essentially provides a way to identify potential interactions between a nonlinear independent effect and time, thus is particularly suitable to characterize independent variable effects that change with time. In clinical applications, it helps to identify and quantify temporal influences of independent variables with nonlinear effects on the longitudinal and survival outcomes, while accounting for the connections between the two outcomes.

Methodologically, the development of the method is no trivial extension of previously published work. The main challenges come from the complex structures of joint models. In a joint model, an independent variable could have completely different temporal effects on the longitudinal and survival outcomes, and thus simultaneously selecting time-invariant and time-varying coefficients in the two model components is technically difficult. Although it is easier to perform model selection for one model component while fixing the structure of the other component, such a piecewise approach fails to take into account of the natural connections between the two components; this may be the primary reason for the lack of development of simultaneous model selection for joint models, especially in the presence of time-varying coefficients. This research uses a penalized likelihood approach, which does not require assumptions of fixing parts of the model. This work demonstrates that simultaneous selection of time-varying coefficients in joint models through penalized likelihood is possible and its implementation is relatively straightforward in complicated modeling situations.

An essential step in the proposed method is the decomposition of the B-spline for independent variable effect into time-invariant and time-varying parts. This decomposition, in practice, is easy to implement. The B-spline basis without intercept could be generated by the R function "bs" with the option "intercept=F", and the computation of the proposed method could be adapted to the existing packages, such as the SAS PROC NLMIXED, thus further extending its applicability. The decomposition may still be improved, as the time-invariant effect is depicted solely by the intercept. When there is limited data information at time zero, the estimation of the intercept, or the time-invariant coefficient may not be sufficiently accurate. In practical data analysis, such as survival analysis, it is not uncommon that most of the observed survival or censoring times are greater than zero. Future work on improving the decomposition may help to improve the estimation of time-invariant effects.

In summary, I showed that by decomposing the B-spline into time-invariant and time-varying parts and then using a penalized likelihood method to select these components, one can identify the time-varying coefficients in joint models. With this in mind, this research has the potential to be used as a practical tool for data analysis.

Chapter 5

Conclusion

Variable selection plays an important role in scientific investigation. To a large extent, the validity of scientific inference depends on the correct specification of the model. In practical data analysis, the analyst has to decide whether a variable should be included in the model, what functional form it should take, and whether the variable interacts with time. In the past several decades, useful model selection procedures have been developed along with necessary selection criteria and statistical tests for most of the commonly encountered statistical models, including linear and proportional hazard models. None of the existing methods, however, are readily applicable to joint models of longitudinal and survival outcomes. The complexity of the joint model has greatly complicated the selection process.

Introduction of LASSO approximately 20 years ago has changed the way in which analysts approach the selection problem. By placing a penalty on model complexity, the method fundamentally simplifies for the selection process. Along this line, various selection methods have been developed for most of the standard statistical models, including generalized linear mixed effects models and proportional hazard models, allowing for selection of both fixed and random effects.

A noticeable gap in the existing literature is the lack of selection procedures for joint statistical models of longitudinal and survival outcomes. The increasing popularity and the widespread use of the joint models present an urgent demand to fill this gap. This dissertation addresses this need in a systematic way, by proposing a series of model selection tools that aid the joint model construction.

In this chapter, I would review the main methodological contribution and practical impact of this research.

First, this dissertation has presented the method to simultaneously select fixed and random effects for the joint model. While the selection of the fixed effects helps to identify independent variables that are related to the outcomes, selection of the random effects serves the dual purpose of specifying the underlying correlation structure and justifying for the joint model structure. Importantly, the reparametrization by Chelosky decomposition allows the random effects in the two model components to retain their own covariance structures, while not restricting the model space by pre-excluding candidate models. Such an approach thus enables researchers to simultaneously perform random effect selection by identifying their corresponding covariance structures. Additionally, the reparametrization also allows the random effects in the longitudinal and survival models to be linked in a common structure. Practically, this reparametrization through Chelosky decomposition has made the selection of random effects by group penalty feasible. This reparametrization and the random effect selection is not restricted to the joint model setting for studying the correlation between the longitudinal and survival outcomes. Actually, it could be extended to any model settings with multiple outcomes to investigate their correlations, which should have wide applicability in clinical investigations.

Second, this thesis developed a method to identify the functional forms of independent variables in an additive joint model. It provides a general semiparametric framework for structural discovery in such a model setting. The decomposition of the B-spline basis clearly partitions the independent variable effect into a parametric (linear) part and a nonparametric (nonlinear) part. I then present the model in a mixed-effect model formulation. Methodologically, the basis decomposition and mixed model representation serve as a bridge between variable selection and structural discovery. Practically, it clearly depicts

the independent variable effects as linear and nonlinear other than lumping them together, thus retaining the model interpretability. The same approach could be similarly extended to other additive models for identifying independent variable effects hidden in the data.

Third, I have also developed a general semiparametric framework to select and estimate the temporal patterns of the independent variable effect. In this framework, a decomposition method is used to partition the temporal effect of a independent variable into time-varying and time-independent parts. This decomposition then allows the application of existing variable selection method to work through the penalized likelihood, eventually distinguishing the temporal effects.

Finally, this dissertation presents a general computational strategy for multiple component models connected by shared random effects. Variable selection and parameter estimation in such models result in intractable integrals. Multivariate Gaussian quadrature method and EM algorithm proposed in this dissertation produce good approximations of the intractable integrals while ensuring computation stability. The deterministic property of the multivariate Gaussian quadrature provides a better way to empirically validate the statistical methods than the Bayesian MCMC approach when dealing with intractable integration, as the latter tends to add another layer of uncertainty during the MCMC sampling. Furthermore, the application of the proposed variable selection and structural discovery method with multivariate Gaussian quadrature is highly adaptable to some widely used existing statistical packages, such as SAS procedure "PROC NLMIXED". This procedure is user friendly and generally suitable to a wide variety of problems.

At the conclusion of this dissertation, I remain hopeful that the applicability of the methods will increase over time. The development of more sophisticated and easier to use packages for implement of the methods will further strengthen the applicability. The methods here are mainly depicted but shall not be limited in the joint model setting.

I anticipate that further modifications and extensions of the current work will become necessary. Future extensions could include variable selections for multiple outcome models and recurrent event failure time models. Dealing with the missing data is an important aspect that I did not study in the current dissertation. Notwithstanding these limitations, I hope that increased application of these procedures will stimulate new thinking for the improvement of the proposed methods.

BIBLIOGRAPHY

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on 19*(6), 716–723.

Albert, P. S. and J. H. Shih (2010). On estimating the relationship between longitudinal measurements and time-to-event data using a simple two-stage procedure. *Biometrics 66*(3), 983–987.

Bondell, H. D., A. Krishna, and S. K. Ghosh (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics 66*(4), 1069–1077.

Brown, E. R. and J. G. Ibrahim (2003). Bayesian approaches to joint cure-rate and longitudinal models with applications to cancer vaccine trials. *Biometrics 59*(3), 686–693.

Choudhry, N. K., W. H. Shrank, R. L. Levin, J. L. Lee, S. A. Jan, M. A. Brookhart, and D. H. Solomon (2009). Measuring concurrent adherence to multiple related medications. *The American journal of managed care 15*(7), 457.

De Gruttola, V. and X. M. Tu (1994). Modelling progression of cd4-lymphocyte count and its relationship to survival time. *Biometrics*, 1003–1014.

Ding, J. and J.-L. Wang (2008). Modeling longitudinal data with nonparametric multiplicative random effects jointly with survival data. *Biometrics 64*(2), 546–556.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *The Annals of statistics 32*(2), 407–499.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Fan, J. and R. Li (2002). Variable selection for cox's proportional hazards model and frailty model. *The Annals of Statistics 30*(1), 74–99.

Fan, J. and R. Li (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association 99*(467), 710–723.

Faucett, C. L. and D. C. Thomas (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: a gibbs sampling approach. *Statistics in Medicine 15*(15), 1663–1685.

Feng, S., R. A. Wolfe, and F. K. Port (2005). Frailty survival model analysis of the national deceased donor kidney transplant dataset using poisson variance structures. *Journal of the American Statistical Association 100*(471).

Friedman, J. H. and W. Stuetzle (1981). Projection pursuit regression. *Journal of the American statistical Association 76*(376), 817–823.

Garcia, R. I., J. G. Ibrahim, and H. Zhu (2010). Variable selection for regression models with missing data. *Statistica Sinica 20*(1), 149.

Hastie, T. and R. Tibshirani (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 757–796.

Hastie, T. J. and R. J. Tibshirani (1990). *Generalized additive models*, Volume 43. CRC Press.

He, Z., W. Tu, S. Wang, H. Fu, and Z. Yu (2014). Simultaneous variable selection for joint models of longitudinal and survival outcomes. *Biometrics*.

Huang, J. et al. (1999). Efficient estimation of the partly linear additive cox model. *The annals of Statistics 27*(5), 1536–1563.

Ibrahim, J. G., H. Chu, and L. M. Chen (2010). Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology 28*(16), 2796–2801.

Ibrahim, J. G., H. Zhu, R. I. Garcia, and R. Guo (2011). Fixed and random effects selection in mixed effects models. *Biometrics 67*(2), 495–503.

Johnson, B. A., D. Lin, and D. Zeng (2008). Penalized estimating functions and variable selection in semiparametric regression models. *Journal of the American Statistical Association 103*(482).

Leng, C. (2009). A simple approach for varying-coefficient model selection. *Journal of Statistical Planning and Inference 139*(7), 2138–2146.

Leng, C. and H. Helen Zhang (2006). Model selection in nonparametric hazard regression. *Nonparametric Statistics 18*(7-8), 417–429.

Lin, Y. and H. H. Zhang (2006). Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics 34*(5), 2272–2297.

Martinussen, T. and T. H. Scheike (1999). A semiparametric additive regression model for longitudinal data. *Biometrika 86*(3), 691–702.

Morrison, L. K., A. Harrison, P. Krishnaswamy, R. Kazanegra, P. Clopton, and A. Maisel (2002). Utility of a rapid b-natriuretic peptide assay in differentiating congestive heart failure from lung disease in patients presenting with dyspnea. *Journal of the American College of Cardiology 39*(2), 202–209.

Nathoo, F. and C. Dean (2008). Spatial multistate transitional models for longitudinal event data. *Biometrics 64*(1), 271–279.

Pinheiro, J. C. and D. M. Bates (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics 4*(1), 12–35.

Pu, W. and X.-F. Niu (2006). Selecting mixed-effects models based on a generalized information criterion. *Journal of multivariate analysis 97*(3), 733–758.

Rizopoulos, D. (2012). Fast fitting of joint models for longitudinal and event time data using a pseudo-adaptive gaussian quadrature rule. *Computational Statistics & Data Analysis 56*(3), 491–501.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics 6*(2), 461–464.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica 7*, 221–242.

Speed, T. (1991). [that blup is a good thing: The estimation of random effects]: Comment. *Statistical Science*, 42–44.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Tsiatis, A., V. Degruttola, and M. Wulfsohn (1995). Modeling the relationship of survival to longitudinal data measured with error. applications to survival and cd4 counts in patients with aids. *Journal of the American Statistical Association 90*(429), 27–37.

Tsiatis, A. A. and M. Davidian (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica 14*(3), 809–834.

Tu, W., G. J. Eckert, T. S. Hannon, H. Liu, L. M. Pratt, M. A. Wagner, L. A. DiMeglio, J. Jung, and J. H. Pratt (2014). Racial differences in sensitivity of blood pressure to aldosterone. *Hypertension 63*(6), 1212–1218.

Wand, M. and J. Ormerod (2008). On semiparametric regression with o'sullivan penalized splines. *Australian & New Zealand Journal of Statistics 50*(2), 179–198.

Wang, H., B. Li, and C. Leng (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(3), 671–683.

Wang, L., H. Li, and J. Z. Huang (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association 103*(484), 1556–1569.

Wang, S., P. X. Song, and J. Zhu (2010). Doubly regularized reml for estimation and selection of fixed and random effects in linear mixed-effects models. *The University of Michigan Department of Biostatistics Working Paper Series*, http://biostats.bepress.com/umichbiostat/paper89.

Wu, L., W. Liu, and X. Hu (2010). Joint inference on hiv viral dynamics and immune suppression in presence of measurement errors. *Biometrics 66*(2), 327–335.

Wulfsohn, M. S. and A. A. Tsiatis (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, 330–339.

Xu, J. and S. L. Zeger (2001a). The evaluation of multiple surrogate endpoints. *Biometrics 57*(1), 81–87.

Xu, J. and S. L. Zeger (2001b). Joint analysis of longitudinal data comprising repeated measures and times to events. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 50*(3), 375–387.

Yan, J. and J. Huang (2012). Model selection for cox models with time-varying coefficients. *Biometrics 68*(2), 419–428.

Ye, W., X. Lin, and J. M. Taylor (2008a). A penalized likelihood approach to joint modeling of longitudinal measurements and time-to-event data. *Statistics and Interface 1*, 33–45.

Ye, W., X. Lin, and J. M. Taylor (2008b). Semiparametric modeling of longitudinal measurements and time-to-event data–a two-stage regression calibration approach. *Biometrics 64*(4), 1238–1246.

Yu, Z., X. Lin, and W. Tu (2012). Semiparametric frailty models for clustered failure time data. *Biometrics 68*(2), 429–436.

Yu, Z., L. Liu, D. M. Bravata, L. S. Williams, and R. S. Tepper (2013). A semiparametric recurrent events model with time-varying coefficients. *Statistics in medicine 32*(6), 1016–1026.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68*(1), 49–67.

Zhang, H. H., G. Cheng, and Y. Liu (2011). Linear or nonlinear? automatic structure discovery for partially linear models. *Journal of the American Statistical Association 106*(495).

Zhang, H. H. and W. Lu (2007). Adaptive lasso for cox's proportional hazards model. *Biometrika 94*(3), 691–703.

Zhangsheng, Y. and L. Liu (2011). A joint model of recurrent events and a terminal event with a nonparametric covariate function. *Statistics in medicine 30*(22), 2683–2695.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association 101*(476), 1418–1429.

CURRICULUM VITAE

Zangdong He

EDUCATION

- Ph.D. in Biostatistics, Indiana University, Indianapolis, IN, 2014 (minor in Epidemiology)
- M.S. in Biochemestry & Molecular Biology, Shanghai Jiao Tong University, Shanghai, 2008
- B.E. in Bioengineering, Shanghai Jiao Tong University, Shanghai, 2005

WORKING EXPERIENCE

- Statistical Intern, GlaxoSmithKline, King of Prussia, PA, 01/2014 - 06/2014
- Research Assistant, Department of Biostatistics, Indiana University School of Medicine, Indianapolis, IN, 08/2010 - 12/2013
- Teaching Assistant, Department of Mathematics, Indiana University Purdue University Indianapolis, Indianapolis, IN, 01/2010 - 06/2010

SELECTED PUBLICATIONS

- He, Z., Tu, W., Wang, S., Fu, H., and Yu, Z. (2014). Simultaneous variable selection for joint models of longitudinal and survival outcomes. *Biometrics*, (In press).
- Li, Y., Tang, K., Zhang, Z., Zhang, M., Zeng, Z., He, Z., He, L., and Wan, C. (2011). Genetic diversity of the apolipoprotein E gene and diabetic nephropathy: a meta-analysis. *Molecular biology reports* **38**, 3243-3252.

- Li, Z., Qu, J., Xu, X., Zhou, X., Zou, H., Wang, N., Li, T., Hu, X., Zhao, Q., Chen, P., et al. (2011). A genome-wide association study reveals association between common variants in an intergenic region of 4q25 and high-grade myopia in the chinese han population. *Human molecular genetics* **20**, 2861-2868.
- Li, Z., Zhang, Z., He, Z., Tang, W., Li, T., Zeng, Z., He, L., and Shi, Y. (2009). A partition- ligation-combination-subdivision EM algorithm for haplotype inference with multiallelic markers: update of the SHEsis (http://analysis.bio-x.cn). *Cell research* **19**, 519-523.
- Zhang, Z., Lindpaintner, K., Che, R., He, Z., Wang, P., Yang, P., Feng, G., He, L., and Shi, Y. (2009). The VAL/MET functional polymorphism in comt confers susceptibility to bipolar disorder: evidence from an association study and a meta-analysis. *Journal of neural transmission* **116**, 1193-1200.
- Ma, G., He, Z., Fang, W., Tang, W., Huang, K., Li, Z., He, G., Xu, Y., Feng, G., Zheng, T., et al. (2008). The Ser9Gly polymorphism of the dopamine D3 receptor gene and risk of schizophrenia: an association study and a large meta-analysis. *Schizophrenia research* **101**, 26-35.
- Li, Z., He, Z., Tang, W., Tang, R., Huang, K., Xu, Z., Xu, Y., Li, L., Li, X., Feng, G., et al. (2008). No genetic association between polymorphisms in the kainate-type glutamate receptor gene, GRIK4, and schizophrenia in the chinese population. *Progress in Neuro-Psychopharmacology and Biological Psychiatry* **32**, 876-880.
- Meng, J., Shi, Y., Zhao, X., Zhou, J., Zheng, Y., Tang, R., Ma, G., Zhu, X., He, Z., Wang, Z., et al. (2008). No significant association between the genetic polymorphisms in the GSK-3β gene and schizophrenia in the chinese population. *Journal of psychiatric research* **42**, 365-370.
- Xu, Z., He, Z., Huang, K., Tang, W., Li, Z., Tang, R., Xu, Y., Feng, G., He, L., and Shi, Y. (2008). No genetic association between ncam1 gene polymorphisms and

schizophrenia in the chinese population. *Progress in Neuro-Psychopharmacology and Biological Psychiatry* **32**, 1633-1636.

- He, Z., Li, Z., Shi, Y., Tang, W., Huang, K., Ma, G., Zhou, J., Meng, J., Li, H., Feng, G., et al. (2007). The PIP5K2A gene and schizophrenia in the chinese population - a case-control study. *Schizophrenia research* **94**, 359-365.
- Ma, G., Shi, Y., Tang, W., He, Z., Huang, K., Li, Z., He, G., Feng, G., Li, H., and He, L. (2007). An association study between the genetic polymorphisms within TBX1 and schizophrenia in the chinese population. *Neuroscience letters* **425**, 146-150.

PRESENTATIONS

- Joint Statistical Meeting, Invited speaker, Boston, MA, 08/2014
- ENAR, Contributed paper, Baltimore, 03/2014
- International Conference on Health Policy Statistics, Contributed paper, Chicago, IL, 10/2013